Varieties of Compatibilism
An Evaluation of Compatibilist Approaches to the Free Will Problem

Wright, John Daniel

Awarding institution:
King's College London

Varieties of Compatibilism:

An Evaluation of Compatibilist Approaches to the Free Will Problem

John Daniel Wright

Student No. 0205613

Ph.D. Thesis in Philosophy

King's College London

Abstract

This thesis critically evaluates some of the main arguments for, and main challenges to, the claim that free will and moral responsibility are compatible with determinism. In the first two chapters I undertake a detailed examination of the so called 'leeway' problem. In chapter one I examine the issue of whether alternative possibilities are required for morally responsible agency by evaluating the Frankfurt example literature. In the second chapter I consider whether the compatibilist can offer an account of alternative possibilities consistent with determinism. Chapter one concludes that the Frankfurt example strategy fails to establish that alternatives are not required. Furthermore, we have good reason to think alternatives are necessary for responsibility. The arguments of chapter two conclude that none of the main compatibilist attempts to capture the alternative possibilities condition are successful.

In chapter three I consider a challenge any compatibilist theory must face independently of the issue of alternative possibilities: the 'source' problem. This is the issue concerning whether we could be the source of our actions in the way necessary for moral responsibility if determinism were true. To this end I will examine Derk Pereboom's four-case manipulation argument in detail here. I conclude that, as things stand, compatibilism is unable to adequately respond to the manipulation argument strategy. Hence we have good reason to think that determinism would mean that we couldn't be the source of our actions in the way required

for free and responsible agency. In summary, compatibilism has not given us satisfactory answers to either the 'leeway' or 'source' worries. Consequently, in chapter 4 I endorse a position in light of this, concluding that although *diagnostic* incompatibilism may be true, we nevertheless can and should be *prescriptive* compatibilists. This is to say, despite the fact that our concepts and practices associated with responsible agency are inconsistent with determinism as they stand, we nevertheless have good reasons to revise and maintain a compatibilist concept of moral responsibility.

Acknowledgements

I first became interested in the free will problem as an undergraduate on the University of London intercollegiate federal BA in Philosophy, as a student at King's College London. Thomas Pink's course on free will and Michael Otsuka's course on ethics with its focus on the Frankfurt example literature sparked an interest that was set to dominate throughout my time as an undergraduate and grad student. I would like to thank them both for starting me off with this topic. I feel particularly lucky that I was in one of the last cohorts to go through the federal BA before it was discontinued. Our lecturers were drawn from across the colleges of the University of London and I took courses under faculty from King's, UCL, LSE, Birkbeck and Heythrop. It was thanks to the federal programme that I was able to attend Otsuka's course as well as Pink's. I have worked on other topics as a graduate student, the mind body problem and contemporary political philosophy in the Rawlsian tradition with the associated problems of global justice, but I've always returned to the free will problem when it came to writing theses and extended pieces of work. It is an utterly gripping problem that once explained demands attention in such a way that makes it hard to disengage with. A great part of our manifest image is challenged by the problem, both metaphysically and ethically in a way that simply can't be ignored. I have now been fortunate enough to be able to write my MPhilStud and PhD theses at King's on free will and responsibility. I would like to thank the UK Arts and

Humanities Research Council (AHRC) for both my masters and doctoral studentships over the last few years without which it would have been impossible to pursue graduate study. The departmental culture at King's College London Philosophy is as much responsible for my choosing to undertake graduate work as my interest in the topics themselves. I will always feel fortunate for having been in such a friendly, inclusive and supportive atmosphere. Thank you to the department!

I am not the first person to say that the free will problem never goes out of philosophical fashion but it is noteworthy that in recent years there has been a profusion of new literature on these issues. I am grateful that my time as a grad student has coincided with this. I would also not be the first to record my debt to the supportive and friendly culture that seems to be the status quo in this sub field of metaphysics. Everyone I've encountered in both Europe and the United States working on free will and responsibility, students and faculty alike, have been a joy to interact and engage with. Faculty going out of their way to help young people working in this field seems to be the norm here.

I would like to thank Maria Alvarez for being my primary supervisor while I've been on the PhD at King's College London. Maria's great insight into these problems and attention to detail with my work have been invaluable and I am very grateful for her support over the last four years. I would also like to thank Bill Brewer for his supervision while Maria was on sabbatical. Bill was very encouraging at a trying time for me and when the writing up pressure was starting to build his advice and sup-

port helped me keep going. John Callanan, my secondary supervisor has also been invaluable. I am grateful to him for both pastoral and intellectual support throughout my time as a graduate student. I would like to thank Tom Pink for letting me have his unpublished material to read and for tutorial support at short notice. I am also grateful to Andrea Sangiovanni who supervised me on the connections between the free will debates and theories of justice early on in my PhD.

I spent the autumn semester of my writing up year in 2013 on the Norman Malcolm fellowship in the philosophy department at Cornell University under Derk Pereboom. I was lucky enough to coincide with Pereboom's excellent graduate course on free will while I was there. I was also supervised by Derk on draft work. His weekly graduate seminars provided an extraordinarily in depth survey of the current state of the literature as well as a pre-publication reading of Derk's *Free Will, Agency, and Meaning in Life.* This helped me greatly in choosing how to structure my thesis. I would like to thank Derk for all his support during my time in Ithaca. I am grateful to Cornell and King's College for funding the Norman Malcolm fellowship and making it possible for me to study in the USA. Michael McKenna has also been a great source of philosophical help and general encouragement over the last few years. Michael has guided me in selecting which topics and issues to work on and put me in touch with other academics at a moments notice when I needed help with recent literature. I am grateful to Michael for his continued support and friendship. I have also benefitted from attending the

London based Action Reading Group. I would like to thank Maria Alvarez, Jennifer Hornsby and John Hyman as well as the other grad students for these productive sessions. Of particular importance to my progress was the close reading the group undertook of Helen Steward's *A Metaphysics for Freedom* in 2012.

I would like to thank my peers for their support and friendship over the last few years: Gary Hayes, Glyn Salton-Cox, Mike Coxhead, Chris Cowie, Patrick Butlin, Peter Sutton, Peter Ridley, Rory O'Connell, Jen Wright, Caspar Wilson, Clare Moriarty, James Arnold, Alex Davies, Alex Geddes, Fay Edwards, Lucy Campbell and Luke Brunning have all helped in countless ways and without them pursuing graduate study would be a different prospect entirely. I am fortunate to be in such a peer group. Lastly I thank my family for everything they have done to support me while I've been a student. Without their understanding of the emotional strains of involvement in this subject I would not have been able to continue.

Contents

Introduction

This thesis is an evaluation of whether compatibilism about free will and moral responsibility can succeed. In this introduction I will start by stating the definitions of free will, moral responsibility and determinism that I'll be working with. I'll then introduce some of the key positions, concepts and general problems around which the free will debate is organised. After that I will outline the overall dialectic of the thesis and summarise the arguments of the individual chapters.

Free Will, Responsibility and Determinism

I understand free will as the control condition on morally responsible action. Free will is the metaphysical control necessary but not sufficient for morally responsible action, whatever that amounts to. On its own free will is not sufficient for morally responsible agency, as there are further necessary conditions as well: epistemic and other psychological conditions such as understanding how your actions can influence others as well as possessing a certain level of moral understanding for example. In defining free will this way I leave it open that there could be value in possessing this type of control that is independent of the link with moral responsibility. For example, we might think it valuable to have free will where this means 'making your life journey your own.' I make no assumptions here other than claiming that free will is the necessary control condition on moral responsibility.

I understand moral responsibility in what has become known as the 'basic desert' sense, following Pereboom:

> This sense of moral responsibility, the one at issue in the free will debate, is set apart by the notion of basic desert (Feinberg 1970; Pereboom 2001, 2007a; G. Strawson 1994; Fischer 2007: 82; Clarke 2005; Scanlon 2013). For an agent to be morally responsible for an action in this sense is for it to be hers in such a way that she would deserve to be blamed if she understood that it was morally wrong, and she would deserve to be praised if she understood that it was morally exemplary. The desert at issue here is basic in the sense that the agent would deserve to be blamed or praised just because she has performed the action, given an understanding of its moral status, and not, for example, merely by virtue of consequentialist or contractualist considerations.[1]

Although I agree with Pereboom that this has been the sense of moral responsibility at issue in the traditional free will debate, at later stages in the thesis I discuss the relevance of other senses of responsibility as well. I do this in the context of chapter four where I examine the potential for concept revision about responsibility.

The determinism at issue in the free will debate is more specifically 'nomological determinism'. As Kadri Vihvelin says:

> we can understand determinism as the thesis that a complete description of the state of the world at any time *t* and a complete statement of the laws of nature together entail every truth about the world at every time later than *t*. Alternatively, and using the language of possible worlds: Determinism is true at a possible world *w* iff the following is true at that world: Any world which has the same laws of nature as *w* and which is exactly like w at any time *t* is exactly like *w* at

[1] Pereboom (2014: 2)

> all times which are future relative to *t*. (See van Inwagen 1983, Earman 1986, Ginet 1990).[2]

Given an initial starting state plus strict laws, the way history unfolds after that point would always be the same even if God 'rewound time' and let things play out again. It's important to be clear that this definition is not committed to a particular analysis of laws and law-likeness. As Vihvelin continues:

> Determinism is a claim about the relation of entailment that holds between, on the one hand, statements of law and statements of particular fact at a time, and, on the other hand, statements of particular fact at any later time. This claim about entailment relations is neutral between different accounts of lawhood, ranging from the so-called "naïve regularity" account (Swartz 1986) to broadly Humean or "best system" accounts (Lewis 1973, Earman 1986, Loewer 1996a, Beebee 2000, Schaffer 2008) to various kinds of necessitarian accounts (Shoemaker 1980, Armstrong 1983, Carroll 1994 and 2008).[3]

With respect to the various arguments and examples discussed in the thesis, I sometimes talk in terms of free will and moral responsibility interchangeably. Given the definition of free will stated above nothing of importance for the question at issue hangs on this.

Incompatibilism

Incompatibilists hold that free will and thus moral responsibility are incompatible with determinism. Under the incompatibilist umbrella there

---

[2] Vihvelin (2011: s1.)

[3] Vihvelin (2011: s1.)

are the 'libertarians' who think we do have free will and hence determinism must be false. There are further distinctions to be made among libertarians in terms of how they understand the indeterministic metaphysics of free will. Some hold that event causation is required, others that agent causation is, and thirdly some think that the powers at work here are non-causal. Yet more subdivisions exist in the form of competing theories of these three families.

There are also the so called 'hard determinists' and 'hard incompatibilists' who are incompatibilists about the concepts of free will and responsibility and also think that we don't have free will. Hard determinists believe determinism is true and hence we don't have free will. Hard incompatibilists might be agnostic about the truth of determinism, or even think it false. But they also think that we couldn't have the kind of free will required for moral responsibility even if determinism were in fact false, or more weakly, that it is hard to see how we could. Both hard determinists and hard incompatibilists can be more or less 'positive' about the fallout for moral responsibility and its associated practices given their positions. For example, at one end of a spectrum, some hard incompatibilists are outright sceptics about moral responsibility and say that nothing remotely like the concept is applicable. In contrast, others hold that although moral responsibility in the basic desert sense itself is not to be had, we can nevertheless make do well enough with something else. Perhaps, for example, a set of reactive attitudes explicitly defined so as to be compatible with determinism (or indeterminism) might be

adopted by a hard incompatibilist. At this other end of the spectrum, a hard incompatibilist might well think that we can carry on much as we were before with our responsibility practices, incompatibilism about the concept notwithstanding.

Compatibilism

Compatibilists hold that free will and determinism are compatible. This might just be because they think determinism poses no threat to the required metaphysics of control or because, more strongly, they think determinism is necessary for free will and believe that we couldn't be in control without it as any indeterminism in a system would introduce randomness or chance in a problematic way. For those compatibilists that don't think determinism necessary for free will, the functionality of their models might not be challenged by indeterminism if, for example, micro level physical indeterminism doesn't 'rise up' or become 'amplified' to effect the functionality of higher level processes involved in agency, or perhaps because, even if that did happen, it wouldn't necessarily impede the requisite control.

Alternative Possibilities

Cutting across the central distinction between incompatibilist and compatibilist and crucial to organising the various positions in the free will problem outlined thus far is the separate distinction between those who think alternative possibilities are required for morally responsible action

and those who deny this. The former are sometimes referred to as *leeway* theorists and the latter as *causal-history* or *source* theorists. It is possible to be either a leeway or a source/causal-history incompatibilist. It is also possible to be a leeway or source/causal-history compatibilist. For example, leeway incompatibilism is defended by Peter van Inwagen and Carl Ginet. Source incompatibilism is defended by Derk Pereboom. Leeway compatibilism (this is sometimes referred to as 'classical compatibilism' in contemporary literature) was defended by A. J Ayer and G.E. Moore and traces its roots back to Hume's discussion in the Treatise. In recent work, Kadri Vihvelin, Michael Fara and Michael Smith have defended sophisticated versions of leeway compatibilism known as the 'new dispositionalism.' Source and causal history compatibilist theories reject the need for alternative possibilities. Frankfurt's own positive 'hierarchical theory' of freedom is one such example. Fisher and Ravizza's 'reasons responsiveness' theory, another.

Challenges for Compatibilism

There are two main challenges that compatibilist theories have to address. These two issues, the *leeway problem*, and the *source problem*, are quickly motivated by the prospect of determinism.

i. The leeway problem

The leeway problem is a problem on the assumption that alternative possibilities are necessary for moral responsibility. If determinism is true, the

thought goes, we would never be able to act otherwise than we in fact do, as there would only be one physically possible future, and this seems to undermine our responsibility since we would be deprived of the necessary alternatives. The leeway problem can be addressed by either showing that the salient kind of alternative possibilities for moral responsibility can still be had in a deterministic world (as the classical compatibilists like Moore and Ayer or the new dispositionalists argue), or by arguing that alternative possibilities are not after all necessary for responsibility and so there is no leeway for determinism to threaten (this is the line taken by the Frankfurt example strategy). The compatibilist burden is to do at least one of these two things.

ii. The source or causal-history problem

Arguing that alternative possibilities are not necessary for moral responsibility does not by itself mean compatibilism is viable. This is because even if it were the case that alternative possibilities were not required for moral responsibility, if determinism is true then our actions would appear to have sufficient causes before we were even born: the previous states of the world and the laws of nature would seem to entail what we do throughout our lives. This gives rise to the worry about source-hood independently of the leeway problem. If determinism were true, the thought goes, then we couldn't be the source of our actions in the way necessary for moral responsibility, even if morally responsible action doesn't require alternative possibilities of any kind. In fact, even for a

compatibilist who defends a leeway position and the necessity of compatibilist alternative possibilities, the source worry is still an issue they need to address as well. Manipulation arguments are not automatically defused because a compatibilist theory has a leeway requirement. That is because, if the manipulation arguments are sound, that kind of leeway won't entail the agent is the source of their actions in the way required if determinism is true.

Thesis Overview

A considerable amount of new material has been published in the last two decades on the free will problem and this thesis will attempt to bring some of the different strands together in an illuminating way in order to get clear on the state of the debate today and the prospects for compatibilism. I will argue that compatibilism has still not managed to overcome its key difficulties, namely the leeway and source problems. However, I will conclude by arguing that despite the failure of compatibilist theories, we should nevertheless adopt a revised concept of moral responsibility that *is* consistent with determinism. In other words, even though I reject *diagnostic* compatibilism I embrace *prescriptive* compatibilism. In summary, I will argue that even though our concept of moral responsibility as it is currently instantiated in ordinary language and practice is inconsistent with determinism, we should revise our concept so it is consistent with determinism.

This may seem like a surprising way to go instead of adopting incompatibilism. However, I am working on the assumption that incompatibilism is not a viable option at this point. I won't defend this claim in the thesis. I'm assuming that incompatibilist theories, whether they are event causal, non-causal or appeal to agent causation are either internally incoherent or, if coherent, unrealistic given our best scientific models of the natural world. The proper evaluation of the various incompatibilist theories to defend this assumption would demand another thesis. Hence, this thesis should be read on the assumption incompatibilism is not a plausible option. I realise this is to bracket half the free will debate but it is necessary in order to give a thorough treatment of the challenges compatibilism faces. There are sophisticated incompatibilist theories in the recent literature, namely Helen Steward's (2012) *A Metaphysics for Freedom* and Mark Balaguer's (2009) *Free Will as an Open Scientific Problem* to name just two central new texts. However, I am persuaded by the arguments against the feasibility of incompatibilism as developed by Pereboom in recent work including his *Free Will, Agency, and Meaning in Life* (2014: Chs 2-3).

Finally, given I'm arguing traditional diagnostic compatibilism fails and also assuming that incompatibilism isn't plausible, I need to justify the move towards revisionism about moral responsibility and prescriptive compatibilism, as opposed to embracing some form of hard incompatibilism or moral responsibility scepticism. Considerations about the methodology of concept revision in general, combined with a normative

argument about needing a concept of moral responsibility, will therefore be developed at the end of the thesis in order to justify revisionism and prescriptive compatibilism in the face of the failure of both the traditional compatibilist and incompatibilist projects.

Summary of Dialectic

The first two chapters will deal with the leeway worry. In chapter one, the question of whether alternative possibilities are necessary for moral responsibility will be addressed. In chapter two, I will evaluate whether compatibilists can provide a satisfactory account of these alternatives if they are required. I will conclude at the end of the first chapter that we have no good reason to believe that alternative possibilities are not necessary for responsibility and then, in chapter two, I will argue that compatibilist attempts to give an account of the necessary alternatives fail. These two chapters alone, if successful, would be sufficient to establish the failure of the diagnostic compatibilist project. However, for the purposes of completeness, given that some of the most prominent compatibilists reject the need for alternative possibilities, in chapter three I will evaluate the manipulation arguments which aim to show that even if (as opposed to what I conclude in chapter one) alternative possibilities were not required for responsibility, determinism would still make it impossible for us to be the source of our actions in the way required for moral responsibility. I will conclude that compatibilists fail to adequately respond to these arguments as well. By the end of chapter three I will have

therefore shown that all the routes to traditional diagnostic compatibilism fail. Given this conclusion and working on the assumption that incompatibilist theories of free will are not coherent or plausible either, I will be left with the job in the fourth and final chapter of justifying the prescriptive compatibilist project instead of embracing hard incompatibilism and moral responsibility scepticism.

Chapter Summary

Chapter 1

In the first chapter I will evaluate whether there is a good argument to establish that alternative possibilities are not required for moral responsibility by considering the Frankfurt-style example strategy. Since Harry Frankfurt published his original case in 'Moral Responsibility and Alternate Possibilities' in 1969, a large literature has developed wherein scenarios are put forward that purport to show agents acting, who are intuitively morally responsible for what they do, yet (due to the structure of the cases), could not have done otherwise. I will consider both the original cases and the full range of modified examples that have been developed over the last four decades. I argue that the strategy fails. I will first evaluate the original examples, sometimes referred to as prior-sign examples, and then go on to look at the no-prior sign cases involving blockage and overdetermination. Finally I will examine the most sophisticated Frankfurt-style cases to date, the so called buffer cases developed

over the last fifteen years. It is within the discussion of a particular criticism of buffer cases, the so called 'timing criticism' that I develop a new line of argument in defence of the principle of alternate possibilities. The timing criticism is one of the most important challenges to the Frankfurt strategy to date but in the recent literature it has come under attack from prominent Frankfurt example defenders. I show that once (i) a key assumption held by both sides in this debate is rejected and (ii) armed with the appropriate understanding of the epistemic requirements on alternative possibility, the principle of alternative possibilities can be defended in a coherent and intuitive way. At the end of the chapter I will also argue that a key new development in the buffer strategy as developed by Pereboom likewise fails to establish that alternatives are not necessary for responsibility. Pereboom's most recent buffer case, *Tax Cut*, as outlined in his 2014 *Free Will, Agency, and Meaning in Life*, is a response to earlier criticisms of his original buffer case *Tax Evasion 2*. I argue that these latest modified buffer cases will not work either.

Chapter 2

Having established in chapter one that we can't conclude alternative possibilities are not required for moral responsibility, and furthermore, that we have good reason to think they are required, in the second chapter I will consider compatibilist attempts to show that the alternative possibilities required are consistent with determinism. The alternative possibilities requirement is traditionally captured by the claim that a

morally responsible agent 'could have done otherwise' than she in fact did. In the first section I will evaluate the so called 'classical conditional analysis' of Hume, Moore and Ayer. This is the thought that to say an agent 'could have done otherwise' is equivalent to saying that an agent 'would have done otherwise, had they wanted (or tried) to do otherwise'. If the conditional analysis were true, these compatibilists (rightly) argued, those conditional statements are consistent with the truth of determinism. However, I will argue that the classical conditional analysis fails as it can be shown that the claimed equivalence doesn't hold by appeal to cases. I'll also discuss a regress worry that threatens the conditional analysis independently of the appeal to cases. In the second section I will go on to examine the so called 'new dispositionalist' analysis of the ability to do otherwise. Recent work on the nature of dispositions by Kadri Vihvelin and Michael Fara has revived the compatibilist project of trying to give an account of the relevant abilities in terms of conditionals, but not the simple counterfactuals of the traditional conditional analysis. The new dispositionalist project is sophisticated but I'll argue that in the end it also fails to capture the sense of 'could have done otherwise' that is at issue in the free will debate.

Lastly, in the third section of this chapter I will discuss recent work by Christian List who has defended a modal interpretation of the ability to do otherwise that explicitly rejects the approach of the classical conditional analysis and the new dispositionalists. Given the initial threat that determinism appears to pose for the ability to do otherwise this might

seem like a counterintuitive move. After all, wasn't the point of trying for a conditional analysis precisely that the unqualified modal analysis looked to be unrealisable if there was only one physically possible future? However, List motivates his argument by drawing a distinction between agentive and physical possibility and claims that restrictions on the latter don't necessarily constrain the former. The methodological approach List uses to defend this move will be examined at length. In the course of my evaluation of List's position I develop two new lines of argument. Firstly, I argue that the sense of agentive possibility List claims is the appropriate one, while sufficient for us *reasoning* over the rationality of alternative possibilities in the way required by rational choice theory, is not necessarily sufficient for the attribution of basic-desert responsibility. In other words, the free will List defends isn't obviously connected up with the concept of moral responsibility in the way it would have to be in order to speak to the traditional problem at issue in the free will debate. Secondly, I provide an independent argument claiming that we cannot divorce agentive modality as List explicates that notion from the micro-physical modalities in the way we'd need to if List's compatibilism is to succeed. Specifically, I develop this argument by considering how the location of the microphysical particles that constitute our bodies in space-time is intuitively relevant to where we, as agents, can be. Hence, contrary to what List claims, what's physically possible for our microphysical constituent particles *is* constraining as to what's agentially pos-

sible for us when we make our choices. I conclude that despite its originality, List's compatibilism is unsuccessful.

Chapter 3

If the first two chapters are sound, I will have shown, firstly, that alternative possibilities are required for moral responsibility and, secondly, that compatibilist accounts of those alternatives fail. That alone would be sufficient to establish the failure of diagnostic compatibilism. However, many prominent contemporary compatibilists, notably Fischer, Ravizza and Sartorio, reject the need for alternative possibilities (because they embrace the Frankfurt-example strategy) and have developed versions of causal history compatibilism. Because of this, in the third chapter I will examine the challenge that manipulation arguments pose for these compatibilists. In other words I will examine compatibilist responses to the source worry independently of the leeway worry. More specifically, I will examine how prominent causal history compatibilists attempt to deal with the most sophisticated manipulation argument in the literature, Pereboom's four-case argument. The point of manipulation arguments such as Pereboom's is to generate a clear intuition of non-responsibility in a case involving deterministic manipulation of an agent and then argue that there is no principled morally relevant difference between such a case and the ordinary deterministic world where no manipulation occurs. Pereboom runs his argument by developing intermediate cases involving less extreme interference between the full blown manipulation

and the ordinary deterministic case to further support the point about there being no morally relevant difference. I will argue that it is hard for the compatibilist to deal with the recalcitrant intuitions of both compatibilist and incompatibilist alike about manipulation cases. It will be necessary to be very careful in reconstructing the dialectic here and to pay special attention to where the burdens lie in these difficult cases. By the end of the third chapter I will therefore have concluded that none of the traditional routes to diagnostic compatibilism succeed, whether they are leeway or source/causal history theories. This conclusion combined with the assumption that incompatibilist theories are not viable leaves me with a choice between hard incompatibilism with some degree of moral responsibility scepticism, or the revisionist route to prescriptive compatibilism. That is the issue that I take up in chapter 4.

Chapter 4

In the fourth and final chapter, having taken stock of the problems that traditional diagnostic compatibilist accounts still face, I will argue for the prescriptive compatibilist project. As I said above, if analysis reveals that we can't square our ordinary understanding of free will and responsibility with determinism, but incompatibilist theories about free will aren't viable, then the fact remains that we may be justified in revising our concepts to make them compatible with determinism. Following Manuel Vargas' work, I build on an argument that proceeds by looking at the methodology of concept revision in other fields and then consider what

parallels can be drawn for moral responsibility. I will argue that there is a good case to be made for the claim that, even allowing for the fact that our existing concept of responsibility has incompatibilist content, enough *other* content constitutes the concept such that we could drop the incompatibilist ideas and still be left with a concept that preserves the majority of the inferential role of moral responsibility. More specifically, we would be justified in continuing to call this a concept of moral responsibility as it would have the majority import of the existing concept. Secondly, and independently of the point about concept revision, I will consider a more direct argument to the effect that we *need* a concept that does at least most of the work of the existing concept of responsibility. This second argument is not necessarily an argument for prescriptive compatibilism per say as a very 'positive' type of hard incompatibilist can run this line of argument as well. Pereboom himself thinks that much of moral practice and value in life as we understand it can make use of and appeal to compatibilist analogues of certain responsibility centred concepts. Given this, I argue it is the combination of these two ideas: the preservation of inferential role on the one hand and the normative argument about the need for a concept of responsibility on the other that constitutes a persuasive defence of prescriptive compatibilism.

Chapter 1

1. Are alternative possibilities necessary for moral responsibility?

1.1 Frankfurt examples

It is very natural to think that in order to be morally responsible for some action an agent must have been able to have done otherwise than they did in fact do. When we hold someone responsible for some action and blame them, we surely think that they should have done something else instead. Furthermore, it wouldn't make much sense to think *that* if they *couldn't* have done something else instead, hence we very easily arrive at the alternative possibilities requirement for moral responsibility. This reasoning is just the 'ought implies can' principle coupled with the thought that the morally blameworthy (and perhaps those rightly subject to the other reactive attitudes as well) ought to have done other than they did do. The conceptual connection was traditionally even considered tight enough to be thought a conceptual truth. However, all of this was challenged when Harry Frankfurt published his paper 'Alternate Possibilities and Moral Responsibility.'[4] In that paper Frankfurt claims to de-

[4] See Frankfurt (1969). Frankfurt credits Robert Nozick with an earlier case with a similar structure. It's important to note that John Locke put forward a related case concerning voluntariness when he considered the man who willingly remains in the locked room. See Locke (1690), *An Essay Concerning Human Understanding*, Bk II, Ch 21 (Of Power), S10.

scribe a case where it seems uncontroversial that the agent involved was morally responsible for what they did and moreover, that they did what they did in the normal way any responsible agent does such things. However, given the structure of the example, Frankfurt also claims that the agent could not have done otherwise. If Frankfurt is right then it would appear we have an example that shows that alternative possibilities for action are not required for moral responsibility. Frankfurt's original example is as follows:

> Suppose someone - Black, let us say - wants Jones to perform a certain action. Black is prepared to go to considerable lengths to get his way, but he prefers to avoid showing his hand unnecessarily. So he waits until Jones is about to make up his mind what to do, and he does nothing unless it is clear to him (Black is an excellent judge of such things) that Jones is going to decide to do something other than what he wants him to do. If it does become clear that Jones is going to decide to do something else, Black takes effective steps to ensure that Jones decides to do, and that he does do, what he wants him to do. Whatever Jones's initial preferences and inclinations, then, Black will have his way.[5]

There are various ways to fill out the details of the example and it's important to fill it out in the most powerful way possible before evaluating its force. So we can assume that Black is able to closely monitor and directly manipulate Jones' brain processes (either directly or remotely) in order to bring about the results Black desires. We should build into the case that Black has the maximum amount of information it's possible to have about Jones regarding his past behaviour and thought processes. That is, he knows the sum total of Jones' previous thoughts, deliberative

[5] Frankfurt (1969: 835)

processes and subsequent actions as well as the resultant patterns and probabilities of future actions given certain deliberative antecedents and counterfactual circumstances. Assume Black knows the lot. The classic rendering of the example then involves Jones deliberating about whether to kill Smith. Black also wants Smith dead for reasons of his own and so waits to see how the situation will pan out without interfering with Jones. As it happens Jones deliberates and decides to kill Smith on his own, in the normal way, without Black having to do anything at all. It is important to note that what actually happens, i.e. the actual sequence of events that leads up to Smith's death at the hands of Jones is *exactly the same* as it would be had Black not been present at all. Black doesn't show his hand and his monitoring doesn't causally interact with Jones so as to affect what he does. However *if* it had looked to Black like Jones was going to do otherwise than kill Smith, then Black would have intervened and manipulated Jones so that Jones kills Smith. The case is presented as one where it's uncontroversial that Jones is morally responsible for killing Smith. Intuitively, Jones is morally responsible for his action and yet, because of the counterfactual process that was never set in motion (by which I mean Black's intervention and manipulation), Frankfurt claims that Jones could not have done otherwise than he did do. Hence we have an apparent refutation of the principle of alternate possibilities.

It's important to note that we must not assume the case takes place in a deterministic world as this would beg the question against the incompatibilist position Frankfurt seeks to refute with the example. Specifically,

it would not be intuitive (at least to the incompatibilist) that Jones was morally responsible for killing Smith if the example was assumed to be deterministic.[6]

1.2 The prior-sign dilemma

It is now widely agreed that Frankfurt's original example fails to coherently describe a case where an agent is responsible without having any alternate possibilities. The original example and cases with analogous structure (sometimes called *prior-sign* examples because Black -or the analogue of Black- has to wait for a sign before they know whether or not to intervene in the process), fall prey to a dilemma. On one horn, Jones (upon closer inspection) does have alternative possibilities; on the other horn, the possibilities are ruled out but so too is the intuition of responsibility along with them. The dilemma is referred to in the literature as the prior-sign dilemma or the Widerker/Kane/Ginet defence.[7] The response focusses on the sign that Black must take as a reason *not* to intervene and let Jones continue to deliberate, choose and act on his own. The thought runs as follows; either that sign not to intervene causally determines Jones' subsequent decision to kill Smith or it doesn't. If it does, then as I said above, the incompatibilist will complain that this is a ques-

---

[6] It should be noted that some commentators, notably John Martin Fischer, disagree that assuming a deterministic setting is dialectically problematic here. I leave this issue to one side. Many incompatibilists would predictably take issue with the assumption of determinism.

[7] See Ginet (1996), Kane (1996), Widerker (1995).

tion begging description of responsible agency and they consequently won't have (or at least can't be expected to have) the intuition that Jones is morally responsible when he goes on to kill Smith. However, if the sign for Black not to intervene doesn't determine Jones' subsequent decision to kill Smith then it seems that Jones has alternative possibilities available to him. He could presumably decide to do something else instead. Even if Black would immediately override that decision (through manipulation of Jones' brain), it nevertheless looks like Jones could have done otherwise, i.e. chosen otherwise. This is consistent with the fact that because of the speed of Black's interventions (when they do occur), Jones doesn't necessarily have alternative possibilities of *macro bodily action* in the example. Specifically, he couldn't necessarily have started to move in some ways that would constitute him following through on a decision not to kill (i.e going over to shake hands with Smith). However, the key point here is just to note that the dilemma defence is not threatened by the fact that there are no alternate possibilities of macro bodily action in the sense of Jones moving around in the context of tables and chairs after a decision not to kill. This is a good way of getting clear on the fact that decision as opposed to macro bodily action is the locus of responsible agency.

Robert Kane and David Widerker were the first to make the type of response outlined above. The very first formulation came in a footnote in Kane's 1985 book *Free Will and Values:*

> ...It is interesting to ask what would happen to a Frankfurt controller if [the action in question were an undetermined choice.] The Frank-

> furt controller would wait to see if the agent would choose A on his own before intervening to make him choose A. But if ... neither the choosing of A nor the doing otherwise [is] determined ... [the] controller cannot tell until the moment of choice itself whether the agent is going to choose A or do otherwise. If the controller wants to ensure that the choice of A is made, he must act *in advance* to bring it about. But if he does this the agent will not be responsible because the agent's choice would have been controlled by the controller. [8]

It might be thought that even though we must allow the case to be indeterministic so as not to beg the question against the incompatibilist, perhaps if Black really was that powerful with regards to his monitoring and knowledge of Jones' behaviour then he would only *not* intervene in those cases where Jones was going to decide to kill Smith on his own. Perhaps Black errs on the side of caution and manipulates in all those situations apart from the ones where it is practically speaking obvious that Jones is going to commit murder. In response to this potential worry it must simply be reiterated that in all those cases where Black doesn't intervene it just *is* possible that Jones can choose otherwise. The structure of the case does nothing to rule that out. Even in a weird possible world where Black and Jones have gone through this process a very large number of times before, and Black is detecting a pattern of deliberation and a weighing of reasons that indicates Jones will decide to kill with a very high probability, Jones might choose otherwise. That's just what it means for a decision to be undetermined. The incompatibilist is entitled to make this claim given their considered position. The only way to get around that would be to make the case deterministic and give Black the ability to

[8] Kane (1985: 51)

calculate infallibly what will happen from past states of the world and the laws of nature. However, as has already been said, this is dialectically unacceptable as the incompatibilist cannot be expected to have the intuition of moral responsibility in that case.[9]

1.3 Problems with Frankfurt's stipulations about the counterfactual case

The prior-sign dilemma outlined above aims to show that, on the only dialectically acceptable version of the Frankfurt case, there is a salient alternative possibility present: namely, when the relation between the sign for Black *not* to intervene and Jones' decision is indeterministic (as it has to be in order not to beg the question against the incompatibilist) it appears that Jones can decide not to kill Smith, notwithstanding the fact that he'll immediately be overruled by Black' intervention.[10] The prior sign dilemma doesn't take issue with the coherence of Frankfurt's case

---

[9] There is some debate about whether it's coherent for libertarians to claim that agents can chose otherwise than they in fact do in very clear cases when all the relevant background desires and reasons strongly favour a particular choice. For example, Steward (2012: 140-144) is an incompatibilist who denies this makes sense in certain kinds of cases. However, this point doesn't undermine my claims above about Jones' decision in the Frankfurt example. It doesn't seen incoherent or puzzling, given Jones' background beliefs and desires, that he should chose not to kill Smith (independently that is, of whether the structure of the Frankfurt case permits such a decision).

[10] When I say sign *not* to intervene here what I strictly mean is the lack of a sign to intervene. The case requires a sign for him to intervene and a lack of a sign for him not to intervene

other than to point out that there is (contrary to what Frankfurt and others have thought) an alternative present.

In addition to the prior-sign dilemma, an independent line of criticism has been levelled at Frankfurt's original case which does challenge the coherence of the stipulations Frankfurt makes about what happens in the counterfactual case where Black intervenes. In the actual case, Jones kills Smith on his own in the normal way and Black's presence and readiness to intervene has no causal bearing on what actually happens. That all seems fine. However, Frankfurt claims that in the counterfactual case where Black intervenes, "Black takes effective steps to ensure that *Jones* decides to do, and that he does do, what he wants him to do. Whatever Jones's initial preferences and inclinations, then, Black will have his way."[11] The worry is that this isn't obviously coherent. More specifically, it isn't clear that in the counterfactual case, the causal chain that Black initiates and which infallibly brings about the required outcome (Smith's death), constitutes a decision *of Jones* at all. It's crucially important to Frankfurt that *Jones* can't avoid killing Smith. But is it legitimate then to stipulate that despite Black's invasive control in the counterfactual case it is still Jones who acts?

Maria Alvarez, Helen Steward and Brendan Larvor articulate this worry. Alvarez, referring to the brain sequence Black sets in motion (in Jones) and also referencing Hornsby's 1980 distinction between transitive (T) and intransitive (I) forms of verbs, writes:

---

[11] Frankfurt (1969: 835)

> Why would making this sequence happen amount to causing *a decision*? Just as the occurrence of a sequence of movements$_I$ of Jones's body is not sufficient for his having performed an action, the occurrence of a particular sequence of neural events is not sufficient for his having made a decision—*even if*, had Jones made a decision, that same sequence would have happened; and even if, whenever Jones makes a decision to perform an action of that kind, a sequence of neural events of the same kind occurs. Thus there seem to be no grounds for thinking that it is legitimate to stipulate, as Frankfurt and others do, that what Black would cause by manipulating Jones's brain would be a decision of Jones's. Indeed, there are compelling reasons for denying that it is legitimate, because none of the criteria for someone to have decided something would apply to Jones in the counterfactual case. For one thing, the 'decision' would not be the outcome of Jones's practical reasoning, or of his emotional response to a situation. But it is in such contexts, i.e. contexts of deliberation, of appraisal and reaction to a situation, etc., that the concept of a decision has application.[12]

Larvor (again referring to Black's intervention in the counterfactual case) writes:

> Black presses his button, initiating a reliable causal chain that ends in Smith's death. This is Black's deed. The fact that the causal chain passes through parts of Jones's body does not make it Jones's deed. Jones is merely Black's unwilling instrument. In this scenario, it would be false to say that Jones kills Smith. Black kills Smith, which is *why* we would hold Black responsible for Smith's death, and not the hapless Jones.[13]

Steward discuses the status of intervention by a manipulative process or agent in the counterfactual case in the context of another modified Frankfurt case (the Mele/Robb case), which I will evaluate later in this

---

[12] Alvarez (2009: 71)

[13] Larvor (2010: 507), my italics.

chapter. In the Mele/Robb case an agent, Bob, is deliberating whether to steal Ann's car. Black wants him to and has in this case initiated a process P in Bob's brain that will ensure Bob decides to steal by some time t only if Bob has not already by that time decided to steal the car. Leaving aside for the moment all discussion of other aspects of this modified case, Steward's comments on this case are equally applicable to the consideration of the status of the counterfactual intervention scenario in the classic Frankfurt case:

> But when process P culminates in Bob's so-called 'decision', Bob seems to have had nothing whatever to do with the resulting event (except that he has, by refraining from thus deciding 'on his own', inadvertently permitted it to occur). The occurrence of this 'decision' in him, because brought about by a process over which he has no control, seems to be just that: an occurrence in him not a decision by him. It was not the consequence, for example, of any deliberation on the part of Bob and presumably bore none of the normal sorts of relationship to such things as his emotional state, the content of prior imaginings, etc. Its causal source, indeed, seems to have had nothing to do with him at all, except that some important parts of the process are located within his skin. Why, then, agree that what has occurred is anything that merits description as a decision, on the part of Bob, to steal Ann's car?[14]

This seems right. It's not at all obvious that interventions to bring about a causal sequence in an agent along these lines could ever constitute a decision and action on the part of that agent. The first thought that comes to mind is that these situations are ones where Black 'acts through' Jones or Bob etc. Further, as Alvarez and Steward both say, the activity that is characteristically associated with a person making a decision

---

[14] Steward (2012: 179-80)

would be absent, i.e. the process Black initiates would not be the outcome of Jones' deliberation and practical reasoning. If this is correct then it would not be the case, contrary to what Frankfurt claims, that this is a situation where Jones is responsible despite the fact that he couldn't have done otherwise. If Alvarez, Steward and Larvor are right that Jones does not act in the counterfactual case, then when Jones kills Smith in the normal way on his own in the actual case, it's not true that *he* could not have failed to act as he does, as in the counterfactual situation, *he* wouldn't act at all. The burden is with the Frankfurt defender to respond to these worries.

Regarding this challenge to the coherence of the counterfactual case in Frankfurt examples, Michael Otsuka and Ezio Di Nucci have discussed the dialectical status (in the debate between compatibilist and incompatibilist) of stipulations about the structure of the counterfactual case. It's important to be sensitive to the burden of explanation and possibility of question begging in articulating these worries. These examples are meant to persuade us that alternate possibilities are not required for moral responsibility without introducing anything contentious. The Frankfurt defender has the task of telling a plausible story which doesn't assume anything that the incompatibilist would take issue with here. Just as it was important that in the actual sequence in the Frankfurt case (where Black doesn't intervene) we assume that determinism is false, we must not assume anything problematic in the counterfactual case either.

Michael Otsuka says the following about the counterfactual case:

> To those who maintain that such neural manipulation is not compatible with agency on the part of Jones, we can imagine, on Frankfurt's behalf, that Black is an omnipotent being who has the power to impose deterministic laws of physics that make it inevitable that Jones kill Smith. Frankfurt's opponent would not want to deny the compatibility of determinism and action, for such a denial would beg the question against Frankfurt, since then, a fortiori, determinism would have to be false for there to be action for which one could be blameworthy.[15]

However, I don't see how in the version of the counterfactual case Otsuka points to here, 'omnipotent' Black avoids the worries articulated by Alvarez, Steward and Larvor. Firstly, it's still the case that after a period of deliberation and authentic mental activity on the part of Jones, when Black intervenes the resultant behaviour of Jones is not connected up in the appropriate kind of way with the previous activity (the activity on the basis of which Black decided he had to intervene) to constitute a decision *of Jones*. Alvarez and Steward's point above seems to retain as much force in this respect however sophisticated Black becomes. The power to impose deterministic laws of physics in such a way that the process that then occurs as a result of this is neurologically identical to the one that would have happened had Jones' deliberation gone differently earlier (more favourably from Black's point of view) still doesn't mean that it is a decision of Jones, rather than Black acting through Jones' neural pathways. The original criticism that there is no decision or action by Jones in the counterfactual case is not itself threatened by Otsuka's point that to assume the falsity of determinism necessary for action itself

---

[15] Otsuka (1998: 687)

would be question begging against Frankfurt here. Determinism was never the source of that worry, it was the disconnection from previous mental activity pre intervention as well as the related worry that causing a sequence of physical events might not constitute causing someone to act at all. The point about disconnection applies equally to both deterministic and indeterministic contexts in these kinds of counterfactual interventions.

The second point of interest is Otsuka's claim that to deny that a deterministic process in Jones (in the counterfactual case) could count as an action here would be to beg the question against Frankfurt. These cases are meant to be stories that persuade people that alternative possibilities are not required for moral responsibility without introducing anything contentious that the target group would reject. That said, why, if it's dialectically unacceptable for a Frankfurt defender to assume determinism in the actual case when Jones acts without intervention (as one horn of the prior-sign dilemma outlined above made clear), is it nevertheless acceptable to assume that Black operates by initiating a deterministic causal process in the counterfactual case? It might look as if what goes for the actual case must also go for the counterfactual case here - but the terrain is more complicated. In the eyes of an incompatibilist about moral responsibility who thought you needed leeway control which is inconsistent with determinism, a Frankfurt defender putting forward a deterministic story in the actual case and expecting an intuition of moral responsibility from the incompatibilist is indeed begging the question. But just

because an incompatibilist is an incompatibilist about *morally responsible* action it doesn't follow that they're an incompatibilist about action *simpliciter*. Plausibly, many people in the leeway incompatibilist camp might conclude that we all still continue to *act* under determinism although no-one is ever morally responsible for what they do. So Otsuka is quite right in this respect. However, for those incompatibilists (like Steward) who are agency incompatibilists, assuming a deterministic story in the counterfactual case is just as dialectically problematic as it is in the normal case (though for different reasons to those outlined here).

Taking stock then, for a large class of incompatibilists, Otsuka is right that assuming Black operates by imposing deterministic causal laws in the counterfactual case is not dialectically problematic *because he's imposing determinism* (though it is for the agency incompatibilists who have independently motivated their positions like Steward). However, Alvarez, Steward and Larvor's original points remain as forceful here *when Black operates as a counterfactual intervener*. The reason for thinking Jones doesn't do anything in the counterfactual case has nothing to do with determinism. It is instead that the neural activity post Black's intervention is not connected up in the right kind of way with the neural activity pre intervention for the latter to plausibly count as an action of Jones. In summary, we have no reason to conclude that Jones is responsible for something he couldn't avoid doing because Jones doesn't *do* anything in the counterfactual case. The history and genesis of an event matter when it comes to whether that event can constitute a decision.

This concludes my discussion of the classic Frankfurt case. In the rest of this chapter these points will be relevant against some of the modified Frankfurt style cases as well.

1.4 What kind of alternative possibilities are relevant in Frankfurt style cases?

The issue of 'robustness'

So far I have only argued that there are alternate possibilities present in a traditional (prior sign style) Frankfurt case. However, in the dialectic with a proponent of any kind of Frankfurt case it's not enough merely to point out that, contrary to the purported structure of the example, alternatives are in fact present. In addition, it has to be shown that these are the right kind of alternatives to ground moral responsibility. More specifically, it has to be the case that alternatives confer the status of responsibility on the agent in the example. In other words, the agent must be morally responsible partly *in virtue of the fact* that they could have done otherwise. In the terms of the debate, the alternatives have to be *robust*. Imagine that a person made an undetermined free choice to do A. The indeterminism means A is not guaranteed, but if (by stipulation) the only other possibilities were involuntary twitches or spasms (things that happened *to* the agent as opposed to voluntary choices or omissions *by* the agent), it's hard to see how the possibility of those spasms or twitches even partly underwrote the responsibility ascription. We don't tend to

think it's in virtue of the existence of those merely possible events that responsibility is conferred. When someone makes a free voluntary choice to do something immoral, the thought is that they didn't have to make that choice and the filling out of 'didn't have to' here should be thicker than the mere possibility that a different event of any kind could have by chance occurred. What we plausibly mean is that something else was not merely possible (in the thin sense of possibility given the indeterminism) but that some other voluntary action or omission was possible, given the deliberative undetermined agency in question. I will now look in more detail at how we should understand the salient notion of robustness here, given recent work by Dana Nelkin and Derk Pereboom.

The central requirement for alternatives to be robust is the demand that the alternative possibilities be exempting. It should be the case that when an agent avails themselves of such an alternative, they would thereby exempt themselves from blame for the thing they would have otherwise done. In Pereboom's words, "The core intuition that underlies the proposal to explain moral responsibility by access to alternative possibilities is that to be blameworthy for an action, an agent must have been able to do something that would have resulted in her being "off the hook."[16] There is an epistemic requirement on robustness: it must also be the case, broadly speaking, that an agent understands that by taking an alternative course to the one they do in fact take, they would have been exempted from the responsibility they in fact have for their chosen path.

---

[16] Pereboom (2014: 10-11)

There are some tricky issues with getting the notion of robustness precisely stated given the nature of the knowledge requirement arising from the need to understand here. In *Living without Free Will* Pereboom initially characterised robustness as follows:

> **Robustness (A)**: For agent to have a robust alternative to her action A, that is, an alternative relevant per se to explaining why she is morally responsible for A, she must have understood that instead she could have voluntarily done something as a result of which she would have been precluded from the moral responsibility she actually has for A.[17]

Pereboom summarises a number of potential problems for this formulation of robustness and then goes on to offer a more nuanced version. To begin with, following a point raised by Jonathan Vance, there is the question of how much credence in the claim that a particular alternative course of action is exempting an agent needs to have. Vance gives the example of an agent who had a belief with a very low degree of subjective probability that a certain course of action might be exempting. Intuitively this would not count as a robust possibility. On the other hand requiring certainty seems far too strong in many cases. This will clearly be a matter of degree for most descriptions of action. Pereboom replies that the threshold credence required will be 'difficult or impossible to determine' but implies that this isn't a special problem for being able to tell what counts as robust. This seems right - each case must be judged individually when it comes to the status of the alternatives an agent has, given their knowledge, beliefs and credences at the time. This is just the general is-

[17] Pereboom (2014: 11)

sue of justification and credence raising its head with respect to the concept of robustness and does not indicate that there is anything wrong with the attempt to characterise robustness this way per say. Secondly, and more importantly, Pereboom discusses a worry Dana Nelkin raises about whether knowledge is even required. Nelkin discusses a case from Mark Twain in which Huckleberry Finn is deciding whether to let Jim the slave boy go free, Pereboom says the following:

> In a familiar example from Mark Twain, Huck Finn sincerely expresses the view that allowing Jim, the slave, to go free, is morally wrong, but nonetheless allows him to go free instead of returning him to his owner. Suppose that he instead, holding that moral psychology fixed, did return him to his owner, and that this is in fact morally wrong. Does Huck have a robust blameworthiness-explaining alternative possibility—i.e., his allowing Jim to go free? My sense is that he might well, despite his not clearly understanding that letting Jim go free is morally right, and that letting him go free would have precluded him from the blameworthiness he actually incurs. For, as Nelkin points out, we suppose that Huck has at least some cognitive sensitivity to the moral rightness of letting Jim go free and the moral wrongness of returning him to his owner. As a result, she thinks that Robustness (A) is too strong, and I believe she is right. Nelkin suggests that understanding isn't needed, but rather only some lower level cognitive sensitivity. I propose that what's required is that Huck has some cognitive sensitivity to the fact that he could do otherwise, and to the fact that if he did do otherwise, he would then be, or would likely be, blameless.[18]

This seems right, requiring full knowledge seems to set the bar too high in cases where we would feel intuitively that the agent had sufficient sensitivity or awareness enough to qualify as a candidate for re-

[18] Pereboom (2014: 12)

sponsibility. Consequently, Pereboom reformulates his robustness condition as follows:

> **Robustness (B)**: For an agent to have a robust alternative to her immoral action A, that is, an alternative relevant per se to explaining why she is blameworthy for performing A, it must be that
>
> (i) she instead could have voluntarily acted or refrained from acting as a result of which she would be blameless, and
>
> (ii) for at least one such exempting acting or refraining, she was cognitively sensitive to the fact that she could so voluntarily act or refrain, and to the fact that if she voluntarily so acted or refrained she would then be, or would likely be, blameless.[19]

In many cases, perhaps most cases in the actual world, agents will have knowledge of, rather than mere cognitive sensitivity to the fact that an alternative is exempting for them. I will henceforth work with this definition (Pereboom's Robustness B) as I examine the different kinds of Frankfurt style cases to determine whether there are in fact robust alternative possibilities present.

Alternatives to do some other thing or just to refrain from what you in fact did?

We need robust alternatives but it seems as if there are two ways of conceiving robust alternatives given the requirements explored in the section above. It's natural to think that whenever an agent makes some decision they are morally responsible for and where they could have done other-

[19] Pereboom (2014: 13)

wise, that means they could have *decided otherwise*. For example, perhaps someone illicitly parks in a disabled parking space when they could have just as easily chosen to park in a regular bay. There are also the cases where the alternative decision is just a decision *to not do the thing you actually did,* though not to do something different instead. For example, I'm responsible for choosing the chocolate cake but I could have chosen *not to take the cake instead* etc. In both the previous examples about parking and cake you choose something but you could have *chosen something else instead*. As the two slightly different examples illustrate, that something else instead might be a different option altogether or just the decision not to do the thing you actually did. It's very natural to offer these kinds of alternatives up as stereotypical examples of what we mean when we are talking about alternate possibilities for action. On the other hand, when we say of someone's decision, 'you didn't have to do that…', we might just mean that they could have refrained, i.e. not done the thing, in a sense that needn't require them to have made some *other concrete decision* at all, whether that other decision is to do something else altogether or just a decision to not do the thing they did do. It looks like there is a weaker and a stronger sense of 'refrain' available in ordinary discourse. You can refrain from choosing A (in the stronger sense) when you make an explicit decision to do something other than A (say B) or by making an explicit decision simply to not choose A. Refraining in the stronger sense is just the same thing as deciding otherwise as illustrated with the examples of parking spaces and cake. However, as already mentioned

above, you can also refrain from choosing A (in the weaker sense) *simply by not choosing A* where this doesn't entail any other voluntary decision (maybe you just continue deliberating). Maria Alvarez argues that the alternative possibilities requirement should be understood in the weaker sense here.[20] Sometimes, elsewhere in the literature this distinction isn't made clear. As we go through the different Frankfurt cases I will make explicit how the incompatibilist is best placed to argue with respect to this distinction. I don't necessarily want to take a firm position on whether the stronger or weaker requirement is the one to endorse here. The most important thing from my perspective is to find (if they exist) robust alternatives in Frankfurt cases. For each type of case evaluated I will discuss the type of alternative present. Either way, if they exist there will be an intuitively grounded robust alternative in a Frankfurt case and that is bad news for the compatibilist project.

In summary, given I have now outlined the importance of both robustness and the distinction between deciding to do something else and refraining from deciding, where both can count as 'doing otherwise', it's clear that the original Frankfurt case has robust alternatives in it. On one horn of the dilemma defence, if the situation is indeterministic and the sign for Black not to intervene doesn't necessitate what happens, when Black doesn't intervene and Jones goes ahead and kills Smith in the normal way, it's true that Jones has the following robust alternatives available. Firstly, Jones could make a decision not to kill Smith (again, this is

---

[20] Alvarez (2009: 63)

consistent with such a decision being immediately overridden by Black's neural intervention). Such an alternative decision would be voluntary and meet the epistemic requirement on robustness. Jones would of course know that by choosing not to kill he was choosing a path that would exempt him from the responsibility he would have incurred for killing. Secondly, Jones could have refrained from deciding to kill Smith where refraining meant nothing more than 'not deciding to kill Smith and not 'deciding not to kill Smith'. Perhaps Jones just continued with his deliberation etc. If this was the alternative then it would also be voluntary and robust as Jones knows that by continuing to deliberate as opposed to deciding to kill he is exempt from the responsibility that would be his had he instead decided to kill. Both alternatives, stronger and weaker, meet the requirements for robustness. I now turn to the evaluation of more recent modified Frankfurt style cases.

1.5 Modified Frankfurt examples

Much of the discussion subsequent to the mid-nineties has revolved around whether it is possible to construct an example that won't fall victim to the prior-sign dilemma. To this end, ingenious new types of examples with different structures have been developed. In addition to the classic prior-sign Frankfurt case there are now no-prior-sign cases involving overdetermination and/or blockage as part of their structure. More recently, buffer cases have been developed which do involve prior signs but in a way that purports to evade the dilemma defence. I will evaluate

each of these different types in turn, taking the best instances of each kind. In addition to possible evasion of the prior-sign dilemma, it will also be important to see if the modified cases avoid the problems associated with the counterfactual case as outlined by Alvarez, Steward and Larvor. One crucial point about example structure is worth keeping in mind here. The principle of alternative possibilities states that in order for an agent to be morally responsible for some action, it must be the case that they could have avoided performing it. In other words, we need a case where there is an instance of unavoidable action, for which the agent in question remains intuitively responsible. That is what you need for a successful Frankfurt-style case. It is consistent with saying *that* that not all alternative possibilities need to be removed from the case structure. For example there might be wiggle room and various ways the agent might act or deliberate in the lead up to the (purported) instance of unavoidable action for which they are intuitively responsible. None of that would matter as long as the case is indeed a case of responsible unavoidable action. Even with loads of other alternatives present the case would still refute the principle of alternative possibilities. It is important to bear this point in mind while exploring the variety of modified cases in the rest of chapter one.[21]

---

[21] I am grateful to Michael Otsuka for making this last point explicit in his comments on an earlier draft of this chapter. In what follows it should be clear how the different types of cases engage (or don't engage) with this requirement. In particular my discussion of buffer cases and the 'timing dialectic' clarifies these issues for the most advanced cases to date.

The Mele-Robb case

In the original prior sign Frankfurt case, when Black doesn't intervene, the indeterminism of Jones' agency left it open that Jones could have chosen otherwise than he did. The ensuring condition (i.e. Black intervening so that the desired 'behaviour' of Jones occurs) did not block off alternate possibilities at the locus of freely willed agency (the moment of choice itself), even if it did mean that at the level of macro bodily action, Jones could never follow through on certain decisions, i.e. by enacting a course of action that Black doesn't want. Is it possible to construct an example where the ensuring condition operates in a way that doesn't leave this space between what the agent actually does (in the normal way on his own) and what the ensuring condition wants to ensure? To this end the Mele/Robb example attempts to describe a situation where two distinct processes each bring about the desired result (the decision to act a certain way) but where the ensuring process does not make use of any prior sign. One process just is the agent deliberating and then deciding in the normal way. We must again assume this first process is indeterministic so as not to beg the question against the incompatibilist. The other process (the ensuring condition) is deterministic and is such that it brings about the desired action by a certain time unless the first indeterministic process itself brings about that same course of action by that time. These processes are independent and when the agent deliberates and chooses as desired in the normal way it's meant to be the case that the deterministic ensuring process does not causally interfere at all in what goes on.

However, if at the crucial moment, the agent fails to choose as desired then the deterministic process would bring about the desired result at just that moment instead. Just like in the original Frankfurt case, if coherent, we would have a situation wherein an agent deliberates and chooses in the normal way and is intuitively morally responsible for what they do yet couldn't have chosen otherwise. The Mele/Robb example runs as follows:

> Our scenario features an agent, Bob, who inhabits a world at which determinism is false... At *t1*, Black initiates a certain deterministic process *P* in Bob's brain with the intention of thereby causing Bob to decide at *t2* (an hour later) to steal Ann's car. The process, which is screened off from Bob's consciousness, will deterministically culminate in Bob's deciding at *t2* to steal Ann's car unless he decides on his own at t2 to steal it or is incapable at *t2* of making a decision (because, for example, he is dead by *t2*) …… The process is in no way sensitive to any 'sign' of what Bob will decide. As it happens, at *t2* Bob decides on his own to steal the car, on the basis of his own indeterministic deliberation about whether to steal it, and his decision has no deterministic cause. But if he had not just then decided on his own to steal it, *P* would have deterministically issued, at *t2*, in his deciding to steal it. Rest assured that *P* in no way influences the indeterministic decision-making process that actually issues in Bob's decision.[22]

What are we to make of this case? Before t2 it is clearly open to Bob to decide not to steal the car and so consequently, any decision *to steal* before t2 is one where Bob has robust alternate possibilities open to him. It is the structure of the situation at t2 itself that is meant to be a counterexample to the principle of alternative possibilities. It seems that Mele and Robb want the following to be true; *P* will *ensure* that Bob chooses to steal

---

[22] Mele and Robb (1998: 101-2)

the car at t2 if he hasn't already and doesn't at that very instant decide himself to steal it. In addition, if he chooses to steal it himself at t2 then *P* doesn't come into it. Is this coherent?

One of the worries here is that it's hard to make sense of how, when x (the agent's own indeterministic decision process) and *P both* coincide on a decision to steal the car at t2, it's x that does the causing and not the deterministic process *P*. Widerker expresses this worry when he says (referring again to the situation where Bob's own indeterministic deliberative process chooses to steal at t2, "what happens in that scenario to the causal efficacy of *P* so that at the time of the decision it is absent?"[23] Widerker calls this the 'Efficacy Problem'. He suggests that this problem may be overcome if:

> …when the two processes are about to culminate in Bob's decision to steal Ann's car, the deterministic process P is somehow preempted. Then, there arises another problem that needs to be resolved: given that P is now preempted and Bob decides to steal Ann's car on his own, there is the distinct possibility that at the last moment he might decide not to steal the car. In that case, the following counterfactual
>
> (**K)** If Bob had *not* decided on his own to steal the car, *P* would have deterministically issued, at t2, in his deciding to steal the car
>
> would be false. The truth of (K), however, is crucial to the success of the Mele/Robb example.".[24]

This is the 'Divergence Problem'. Mele and Robb attempt to deal with these issues by adding more detail to their example:

---

[23] Widerker (2003: 54)

[24] Widerker (2003: 55)

> [There are] two different 'decision nodes' in Bob's brain. The 'lighting up' of node N1 represents his deciding to steal the car, and the 'lighting up of node N2' represents his deciding not to steal the car. Under normal circumstances and in the absence of preemption, a process's hitting a decision node in Bob 'light's up' that node. If it were to be the case both that P hits N1 at t2 and that [the indeterministic process] x does not hit N1 at t2, then P would light up N1. If both processes were to hit N1 at t2, Bob's indeterministic deliberative process, x, would light up N1 and P would not.... [Furthermore], although if both processes were to hit N1 at t2, Bob's indeterministic process, x, would preempt P and light up N1, it is also the case that if, at t2, P were to hit N1 and x were to hit N2, P would prevail. In the latter case, P would light up N1 and the indeterministic process would not light up N2. Of course, readers would like a story about why it is that although x would preempt P in the former situation, P would prevail over x in the latter. Here is one story. By t2, P has 'neutralized' N2 (but without affecting what goes on in x). That is why, if x were to hit N2 at t2, N2 would not light up. More fully, by t2, P has neutralized all of the nodes in Bob for decisions that are contrary to a decision at t2 to steal Ann's car (for example, a decision at t2 never to steal anything). In convenient short hand, by t2, P has neutralized N2 and all its 'cognate decision nodes.[25]

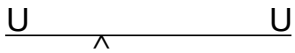
Even granting all this, it still seems the elaboration leaves at least Widerker's 'efficacy' worry unresolved. Mele and Robb have given us a story which explains how x cannot light up N2 or other nodes at t2 because they have (by then) been neutralised *but the issue of how when x activates N1, P doesn't do the causing is left unanswered*. In other places however, Mele and Robb do directly engage with the efficacy problem when they discuss the analogy of a widget making machine and also make appeal to models using seesaw-devices and ball bearings.[26] Their defence of

---

[25] Mele and Robb (1998: 105)

[26] Mele and Robb (2003: 132-6)

(the conceptual possibility of) *current preemption* appeals to a see-saw device like the one below.



The machine is said to have right bias in the sense that when two ball bearings hit the cups at the same time the right cup will be pushed down and not the left come. Despite the ball bearings reaching the cups at exactly the same time, we can say that the rightmost one caused the relevant change. Applied to the Mele/Robb case, the idea is that the deterministic process would terminate in the cup without the bias and the indeterministic deliberation terminates in the cup with the bias. Hence we can make sense of saying (as Mele and Robb want to) that the indeterministic process does the causing and not the process P, even though they hit N1 at the same time.

However, even granting that this ball bearing machine and see-saw analogy make occurrent preemption (trumping) conceptually possible for the sake of argument, there is still a problem for the Mele/Robb case. It's the issue of the status of the causal process that is put in place when P does the causing and not x at t2. Following Otsuka's point from the discussion of the counterfactual case in the classic Frankfurt example, I won't take issue here with the fact that the process P is deterministic. The problem is that when it does the causing, process P is subject to exactly the same worry that Alvarez, Steward and Larvor articulate. In the

Mele/Robb case, process P starts a causal chain in Bob's brain which is suitably disconnected from Bob's previous neural activity before t2 that 'P lighting up N1 at t2' when it occurs, is not a candidate for being *constitutive of an action of Bob's*. Just as the neural events produced as a result of Black's intervention in the classic case are not appropriately the outcome of Jones' deliberation or the result of his emotional response to the situation with respect to whether or not to kill Smith, analogously, we can counter Mele and Robb by saying that Process P is not appropriately the outcome of Bob's deliberation or the result of his emotional response to the situation involving Anne's car. The same worry holds. To suddenly insist that P lighting up N1 is a decision of Bob's is to stipulate something implausible given the uncontroversial historical conditions on decisions. As Larvor might say, when P lights up N1 "it's Black's deed, not Bob's" etc. Consequently, Mele and Robb have not coherently described a case where Bob is responsible for something he couldn't avoid doing as in the counterfactual case *he* doesn't act. The principle of alternate possibilities is not threatened by the Mele/Robb case for the same reason here: the worry about the coherence of the purported agency in the counterfactual case.

Pure blockage cases

The classic blockage case in the literature was developed by David Hunt.[27] As is the case with other Frankfurt style examples, the aim is to describe a situation where an agent freely chooses a course of action in the usual way and where we find them intuitively responsible. However, given the structure of the case, they couldn't have done otherwise. No sign and counterfactual intervention strategy is made use of by Hunt and instead it is stipulated that all of the neural pathways that constitute the supervenience bases of alternative decisions are blocked off or neutralised. The Mele/Robb example just examined is (in part) a blockage example as it was stipulated that other decision nodes are neutralised in Bob. I didn't take issue with the blockage part of the Mele/Robb example as the problem with the counterfactual case when process P activates N1 was sufficient to establish that Mele and Robb's story doesn't work, although it is interesting to consider how similar the Mele/Robb example would be to Hunt's blockage case if modified to exclude the process P and retain simply the 'neutralised other decision nodes' part of the story. However, Hunt's example doesn't involve any other mechanism apart from blockage. After considering John Locke's discussion of the man who willingly stays in a locked room, Hunt develops an example in the following way:

> What is needed is a case in which unavoidability arises solely from blocked alternatives (rather than compulsion in the actual sequence),

---

[27] Hunt (2000: 195-227)

> and the amount of elbow room remaining to the agent approaches zero. Imagine then a mechanism that blocks neural pathways rather than doorways. Suppose that the actual series of Jones's mental states leading up to the murder of Smith is compatible with PAP, except that the mechanism is in operation. The mechanism is not intervening directly in the series itself; it is allowing the series to unfold on its own, but simply blocking all alternatives to the series. Of course it can't block alternatives *in response* to the way the series is unfolding, because then the blockage would be coming too late to have any effect on the avoidability or unavoidability of Jones's actions. Instead, the mechanism blocks alternatives in advance, but owing to a fantastic coincidence the pathways it blocks just happen to be all the ones that will be unactualized in any case, while the single pathway that remains unblocked is precisely the route the man's thoughts would be following anyway (if all neutral pathways were unblocked). Under these conditions, the man appears to remain responsible for his thoughts and actions, …[28]

Granting for the moment that the case is coherent, we can still legitimately ask why the agent in question couldn't have refrained from deciding in the way they did in fact decide. In line with the distinction made above between making another decision on the one hand and simply refraining from deciding as you did on the other, we can ask why the weaker sense of refraining here should also have been ruled out in this case. If the blocked neural pathways are meant to be the supervenience bases for other possible decisions then maybe Hunt is indeed describing a situation where Jones couldn't make any other decision apart from the one he does in fact make but that doesn't entail that Jones could not have simply refrained from making the decision he does make. As Alvarez says, refraining from doing x need not amount to anything more than

---

[28] Hunt (2000: 218)

simply failing to do x here: provided certain other conditions are met, the refraining will meet the condition for robustness. The progress down the only neural pathway available wasn't inevitable because the situation is indeterministic. As with all Frankfurt style cases the actual sequence (though as we saw above not necessarily the counterfactual sequence) has to be indeterministic in order not to beg the question against the incompatibilist. Given that, and the structure of the set up, it seems perfectly reasonable to claim that Jones could have refrained where this would simply correspond to his neural activity not going down the only available path given his awareness of what's at stake. It seems hard to see how Hunt can avoid this possibility given the dialectical constraints on the example here.[29]

The reply above which Alvarez presses is sufficient to conclude that there are robust alternative possibilities in a blockage case as the agent can still refrain. However, it is interesting to consider whether there might be other alternate possibilities in blockage cases as well as the possibility of simply refraining from deciding. The blockage strategy has been described as an attempt to construct an 'in the head' analogue of John Locke's example of the man who willingly chooses to remain in a locked room.[30] If so, just as we might wonder whether the person in Locke's room could have tried the door handle, we might now ask

---

[29] Alvarez (2009: 69, fn. 12) develops exactly this point against the blockage strategy

[30] See Locke (1690), *An Essay Concerning Human Understanding*, Bk II, Ch 21, S10.

whether Jones in Hunt's case could have 'tried' a blocked neural pathway. Is this a coherent possibility given Hunt's stipulations about the blockage case? The story we are told by Hunt is a physiological story about brain structure. Getting on board with this example and taking it seriously requires a certain level of commitment to a scientific supervenience based model of the mind but it's important to tread carefully when considering how the electrical signals in the brain might behave, and consequently what that entails regarding the possibilities for the agent in this context. We know we can't assume the process is deterministic because of the dialectical situation the Frankfurt example defender is in with the incompatibilist. Hunt wants to block off any neural pathways that weren't going to be used anyway but we might still wonder whether neural activity that could plausibly be construed as the supervenience base of 'trying to decide otherwise' can be so uncontroversially ruled out here. In short it's not clear to me that given the indeterminism, electrical activity might not 'flicker about' in such a way that it *would have* activated another neuronal pathway (the supervenience base for an alternative decision) had Hunt's blockage stipulations not been in place. I don't need to be too specific here about the physiological story, the dialectical burden means that's Hunt's job. It is sufficient if I point to something plausible that might constitute the analogue of John Locke's man trying the handle on the locked door. If such indeterministic veers of activity (towards blockages) were possible here then I think there is an ar-

gument to be made that they would constitute robust alternate *decisions*. I develop this argument below.

Carl Ginet has argued elsewhere that decisions are discrete events. If you've resolved to take a certain decision then there is no temporal interval which needs to be completed from start to finish before we can say that the decision event is complete.[31] I want to use Ginet's point about the discrete status of choice or decision events in combination with a clarification of the proper status of these possible 'electrical veers' to argue that there might be robust alternate decisions in Hunt's blockage case. This is especially important given the worries John Martin Fischer has articulated about flickers of freedom not being robust enough in the context of Frankfurt style scenarios.

Might not electrical veers towards other (blocked) nodes constitute decisions to do otherwise here? Given the discrete structure of decisions it seems plausible that if we have the setting out towards a blocked node then we arguably already have a decision. We are considering an indeterministic context, where an agent is deliberating what to do and is fully aware of the different options and knows nothing of Black or any blockages.[32] More specifically, if (i) it was true that the indeterminism crucial for the incompatibilist here was in the right place (i.e. between the delib-

---

[31] Ginet (1996)

[32] I am admittedly shifting here between talk of the neural behaviour of electricity and the behaviour of the agent when I talk about veers. I am assuming that the neural supervenience base of an agent's decision could be such an electrical veer here.

eration and the initiation of the veer) and in addition (ii), if it was true that had the veer been allowed to develop *without* Black/Process P/ Hunt's blockages it would have (in all close possible worlds without interference) been continuous with the causal processes to translate such a decision into the relevant set of bodily actions, *then it seems correct to say that a decision has already been made.* These two conditions are not ruled out by any part of the blockage case structure and together they constitute an argument for the claim that we can understand the veer itself as a decision and that we don't need to worry that nodes can't be activated. If I'm correct then the nodes are nothing special, they're just another part of the downstream neural journey after the actual decision (the veer) has already been made. It's difficult to see this at first given how Hunt describes the case (Mele and Robb do the same with respect to their blockage stipulations). It's stipulated that other decision nodes have been neutralised in these cases and hence it looks to be begging the question by pressing this particular line. However, on reflection, it appears acceptable to push this worry given the scientific framework Hunt is telling his story in and moreover, understanding the status of the veers in the way outlined above seems in line with what we would expect from a leeway incompatibilist model and therefore not *ad hoc*. It's up to Hunt to refine the physiological story in such a way that the move I make here is ruled out. Moreover, if the veers towards Hunt's blocked alternatives do constitute full decisions to do otherwise then they uncontroversially meet the requirements for robustness.

Perhaps it might be said in reply to the argument above that if the other neural pathways are neutralised then you can't even start a veer in the way I have suggested. There would be no wiggle room between the actual sequence of deliberation and decision as it unfolds and any any other neural pattern whatsoever. As I have already said however, given the indeterminism that must be present in these cases it isn't clear one can at the same time simply stipulate this. In the end though, even granting the coherence of Hunt's case and the impossibility of veering, Alvarez's point about there being robust alternate possibilities of refraining in blockage cases retains its force here.

At this point in the dialectic, it's important to remind ourselves of the methodology in use and its parameters. The task of the Frankfurt defender is to tell a story in which we find an agent intuitively morally responsible in the normal way for some action that was unavoidable. So far the stories we have been considering have been mostly about the human brain and manipulated neural structure, this was true of the original Frankfurt case and then again with Mele/Robb and Hunt. In the above sections I have argued that the leeway compatibilist can still consistently find robust alternative possibilities within those cases. Either in the (weaker) sense of having the ability to refrain from deciding what was in fact decided, or the (stronger) sense of having fully blown alternative decisions available. Having evaluated these classic cases and found them wanting I now turn to the latest and most sophisticated Frankfurt style examples - the 'buffer cases'. Interestingly, the buffer cases have not fo-

cussed on trying to rule out alternative possibilities per say but instead on specifically trying to ensure there are no robust possibilities present at all.

1.6 Buffer cases

The two main buffer examples in the literature are Pereboom's *Tax Evasion 2* and Hunt's *Revenge.* Recently, Pereboom has also developed a new modified case, *Tax Cut*.[33] As with the classic counterfactual intervener Frankfurt cases, the buffer cases also make use of a sign to intervene but the difference is that the sign in a buffer case is meant to be a necessary but not sufficient condition for the agent availing themselves of a robust alternative.[34] The sign to intervene (the necessary condition) is the buffer that needs to be crossed before robust alternatives become available to the agent, but the cases are designed such that, allegedly, crossing the buffer is not *in itself* a robust alternative. What crossing the buffer does, is immediately initiate the intervention and manipulation as in the standard prior-sign Frankfurt style cases. I will outline Pereboom's *Tax*

---

33 See Pereboom (2000, 2001: Ch 1.) and Hunt (2000, 2005) for the original formulations of the *Tax Evasion* and *Revenge* cases. See Pereboom (2014: Ch 1.) for *Tax Cut* and his up to date position on the buffer strategy

34 By 'not sufficient' here I simply mean that it is consistent with crossing the buffer that the agent still not decide in the way the buffer enables them to decide. In one sense crossing the buffer is sufficient for the ability to decide that way (that they previously couldn't before they crossed the buffer), but it's not sufficient for that choice simpliciter, as it's still up to the libertarian agent which way they decide.

Evasion 2 below and argue that the principle of alternate possibilities is not challenged. Pereboom's case runs as follows:

> ***Tax Evasion 2***: Joe is considering claiming a tax deduction for the registration fee that he paid when he bought a house. He knows that claiming this deduction is illegal, but that he probably won't be caught, and that if he were, he could convincingly plead ignorance. Suppose he has a strong but not always overriding desire to advance his self-interest regardless of its cost to others and even if it involves illegal activity. In addition, the only way that in this situation he could fail to choose to evade taxes is for moral reasons, of which he is aware. He could not, for example, fail to choose to evade taxes for no reason or simply on a whim. Moreover, it is causally necessary for his failing to choose to evade taxes in this situation that he attain a certain level of attentiveness to moral reasons. Joe can secure this level of attentiveness voluntarily. However, his attaining this level of attentiveness is not causally sufficient for his failing to choose to evade taxes. If he were to attain this level of attentiveness, he could, exercising his libertarian free will, either choose to evade taxes or refrain from so choosing (without the intervener's device in place). However, to ensure that he will choose to evade taxes, a neuroscientist has, unbeknownst to Joe, implanted a device in his brain, which, were it to sense the requisite level of attentiveness, would electronically stimulate the right neural centers so as to inevitably result in his making this choice. As it happens, Joe does not attain this level of attentiveness to his moral reasons, and he chooses to evade taxes on his own, while the device remains idle.[35]

The salient feature of the case is allegedly that although there are clearly alternate possibilities, no *robust* alternative is present. Joe *could* have voluntarily of his own libertarian free will have been more attentive to moral reasons instead of not being so attentive. However, according to Pereboom this is not a robust alternative because, he claims, without the device in place, being more attentive is still consistent with Joe either

---

[35] Pereboom (2014: 15)

choosing to evade taxes or choosing not to. Joe therefore wouldn't know that by choosing to be more attentive he would thereby be avoiding responsibility for choosing to cheat because he would still see whether he chooses to cheat as undecided and up to him (even though with the device in place, he would in fact be immediately 'forced' to evade taxes and thus in actual fact avoid responsibility for choosing of his own accord - he doesn't know that). Being more attentive (crossing the buffer) is the necessary condition for Joe being able to actually *choose not to* evade taxes. It's not a sufficient condition because it is consistent with crossing the buffer that Joe nevertheless still chooses to cheat. Consequently, crossing the buffer doesn't satisfy the epistemic requirements on robustness according to Pereboom.

It's also instructive to consider the other main buffer case in the recent literature before evaluating the strategy in general. David Hunt asks us to consider the following case, called *Revenge*:

> ***Revenge***: Jones, while attending a party, is deliberately humiliated by Smith. The first thought that occurs to Jones, after realizing what Smith has done to him, is that he would like to kill Smith. He leaves the party, escaping the immediate pressures of the situation and giving himself ample opportunity to pull back from this line of thinking. Given the kind of person Jones is, and given the situation in which he finds himself, the alternative of not killing Smith is not unthinkable for him; moreover, should he entertain this alternative, nothing would prevent him from deciding and acting on it. But Jones could decide (and act) otherwise only if he first considered acting otherwise, and he never does this (though he could); instead, he nurses his grievance without respite, while the idea of killing Smith becomes more and more attractive to him. Having finally decided to do the deed, he gets a gun from his car, returns to the party, and shoots Smith dead …. The final element to be added to Revenge is the coun-

> terfactual intervener, which differs from the device in … [earlier FSCs] … inasmuch as it is programmed to hijack Jones's mental processes and force him to decide to kill Smith if he so much as considers not killing Smith. With this device in place, there is no alternative to Jones's deciding to kill Smith: Jones can decide otherwise only if he first considers doing so, but then the device will force him to decide to kill Smith. So an alternative decision is not available to Jones in Revenge...[y]et Jones, who in fact proceeds to murder Smith on his own, leaving the device un-triggered, seems morally responsible for killing Smith.[36]

As with Pereboom's *Tax Evasion 2*, the agent in *Revenge* uncontroversially has access to alternative possibilities. The aim of the case is to show that none of them is robust and therefore they can't ground the intuitive attribution of moral responsibility. Jones in *Revenge* can be granted whatever powers libertarian free will requires. The crux is in the stipulation, 'But Jones could decide (and act) otherwise only if he first *considered* acting otherwise, and he never does this (though he could).'[37] The counterfactual intervener only acts if the necessary condition for doing otherwise is instantiated, that is to say if the buffer is crossed.

1.7 Criticism of the buffer strategy

Are the buffer stipulations coherent with our concept of libertarian free will?

I will explore a number of potential problems with the buffer strategy below. I shall first examine whether Pereboom is justified in stipulating

---

[36] Hunt (2005: 132–134)

[37] Hunt (2005: 132)

the buffer as he does in *Tax Evasion 2*, before going on to examine internal problems, granting the example set up as Pereboom wishes.

Upon initially encountering this case one might worry that it wasn't legitimate to stipulate Joe had to cross the buffer of raising his moral attentiveness before he could access the possibility of making a contrary decision. You might think this for two reasons, which taken together seem to suggest otherwise. Firstly, Pereboom allows that Joe has libertarian free will in the example and secondly, the epistemic condition on responsibility *is* met when Joe goes ahead and voluntarily chooses to evade taxes 'on his own' without any intervention occurring. So although Joe allegedly doesn't have sufficient moral attentiveness in order to be able to robustly choose otherwise, he must have sufficient moral attentiveness to meet the epistemic condition on responsibility. The latter means he has sufficient knowledge of the moral status of tax evasion and its implications to be held blameworthy when he chooses to evade. However, it would seem that if you had that level of knowledge and libertarian free will then there would be nothing to stop you simply choosing not to evade at any point. If you knew that much and your will is free, why can't you simply choose otherwise? It would appear to be a part of our folk understanding of free will that you don't necessarily need full moral attentiveness (or awareness) with respect to a course of action in order for it to be a live option for you, so to speak. Joe has sufficient understanding of the status of his actions to be blameworthy so it seems this level of moral sensitivity should be sufficient for Joe to be able to try to

avoid this course. This thought might be put in the form of the following dilemma. Either Joe has sufficient moral understanding of tax evasion to be blameworthy or not. If he doesn't, then he can't be responsible when he evades and we don't have a candidate for a Frankfurt case. If on the other hand Joe does have sufficient moral understanding to be blameworthy, then regardless of whether he's actually considering these issues as he thinks things over, it looks like that same level of understanding would also be enough to make not evading a live option for him (contrary to what Pereboom claims). I agree that even with libertarian metaphysics in the example there are many decisions Joe just can't make. For example, perhaps it's true to say that he couldn't just decide to pursue a course of action that had never entered his mind and that he had no knowledge or understanding of. But given that Joe does have relevant moral attentiveness here, coupled with the (allowed for) libertarian freedom, it would appear that the stipulation of the buffer is in tension with these features of the case.

Following this line of thought, it seems right to say that If you knew you shouldn't do something and you had libertarian free will and you were not subject to coercion or manipulation or irresistible desires etc then you can simply decide or attempt (when the time comes) to not do that thing. That surely falls out of the folk idea of libertarian freedom if anything does. It is consistent with saying this that some choices are both impossible or very difficult to make. Firstly, as I said above, when you don't know about an option at all, it seems right to say you just can't

choose it. Secondly, there are cases where we know certain courses of action are very frightening, dangerous and difficult. There are situations where, though we know what we should do, it's very difficult to do it for other conflicting moral reasons, but we don't think we can't do these things in the sense that's at issue here, i.e. that they are impossible. In such cases of great difficulty, depending on the specifics, we may attribute diminished responsibility to the agent in question, but we don't do this because they couldn't do otherwise. We think their responsibility is mitigated because it was a lot harder to do otherwise than it would otherwise have been and we empathise with this given what we realistically expect from people. However, despite the possibility and coherence of these kinds of cases, Joe's situation in *Tax Evasion 2* is neither in the former nor the latter category. In summary then, it's hard to see how the stipulation of this buffer is legitimate given that Pereboom allows libertarian freedom and sufficient moral awareness in the case.

However, this worry about the legitimacy of stipulating the buffer misses the mark here. Even if everything I have said above is true of how *our folk concept* of libertarian free will is ordinarily supposed to work, Pereboom can just ask us to imagine a world as close to our own as possible but with the buffers stipulated as described. This is a legitimate move and does not beg the question against the libertarian. Even if it's true that ordinarily, in our closest possible libertarian worlds, the moral attentiveness Joe possesses before crossing the buffer would be sufficient to allow Joe to simply choose otherwise, there is logical space for Pere-

boom to ask us to consider his buffer case. In short, there is nothing contradictory about the buffer stipulation even if, in the closest libertarian world, Joe would *not* need to cross a buffer in order to choose not to evade. We should therefore imagine it's simply true by fiat that the buffers operate as described. In the example Joe deliberates and then chooses in exactly the same way he would have done had the buffer not been there. Intuitively, it seems he is morally responsible for choosing to evade taxes just as he would be in the normal case without the buffer and intervener. *If* he doesn't have any robust alternatives then this example would undermine the principle of alternate possibilities.

As a different rejoinder to this kind of buffer example, perhaps it might be suggested that our intuition of responsibility isn't uncontroversial with these buffers stipulated. Specifically, if we're asked to imagine Joe *with* libertarian freedom, it might be that (given the normal inferential import of libertarian freedom as discussed above) we are still tacitly imagining him with alternatives and hence our intuition is not reliable precisely because he is not meant to have any here. On reflection though, I feel I am able to clearly consider the case on the assumption that Joe doesn't have robust alternatives present by focussing carefully on the structure of the example. Joe chooses to evade in the same way he would in a libertarian possible world without the buffers. Furthermore, Joe remains intuitively morally responsible. I don't think therefore that there is anything inherently problematic with the stipulation of the buffers themselves. These last considerations are important as they make clear exactly

what is being argued about here. At issue is the idea that robust alternate possibilities for action are *always required* for moral responsibility. The leeway incompatibilist is claiming that it's a necessary condition on the concept of responsibility that such alternatives are present. The point to note is that it's consistent with the falsity of this last claim that our *actual world* concept of libertarian freedom could be in tension with the stipulation of the buffers in the way I outline above. Again though, even if this were true, if on reflection we can coherently conceive of a case where an agent chooses voluntarily in a buffered scenario and is intuitively responsible, we might have an argument for the claim that robust alternatives are not necessary for responsibility which is the claim at issue in this debate.

Derivative or direct responsibility in buffer cases?

In the literature, some PAP defenders have agreed that there are no robust alternative present in the buffer cases but have countered that this is because the intuition of moral responsibility in these cases tracks derivative and not direct responsibility. They go on to argue that PAP applies only to cases of direct responsibility. They claim that there will be some prior decision for which there were robust alternatives. For example, Joe's decision in *tax evasion 2* can be traced back to prior failures to make better choices (in situations where PAP applied) it will be countered. Hence buffer cases do not show robust alternate possibilities are

not required for (direct) moral responsibility, which is what they take to be at issue in the debate.

One classic example of derivative moral responsibility mentioned by Pereboom is the case of getting drunk when you know you'll be unable to manage your temper while intoxicated. Despite this, you go on to get drunk and end up getting into an argument and assaulting one of your companions. At the time of the assault, standard conditions on direct moral responsibility are not met. You are not able to sufficiently control your emotions and rationalise courses of action and so you are not directly responsible for the assault when it occurs. However, because you got drunk knowing you have this problem and knowing you would be coming into contact with people, the decision to get drunk, or the failing to choose not to get drunk (for which we can assume there were robust alternatives available) is one for which you are directly responsible. This makes the subsequent assault something you are responsible for derivatively. Another classic example is the case of drunk driving when you hit and kill a child 'under the influence'. Could something analogous be going on with the buffer cases? Specifically, could the agents in buffer cases be derivatively morally responsible for their blameworthy actions in light of earlier choices and omissions for which they had robust alternatives courses of action? Widerker and Ginet have both put forward this worry. Widerker says:

> [A] problem with Pereboom's example is that, in it, the agent is *derivatively* blameworthy for the decision he made, because he has not done his reasonable best (or has not made a reasonable effort) to

> avoid making it. He should have been more attentive to the moral reasons than he in fact was—something he could have done. And in that case, he would not be blameworthy for deciding to evade taxes, as then he would be forced by the neuroscientist so to decide. If this is correct, then Pereboom's example is a case of derivative culpability, and hence is irrelevant to PAP, which … concerns itself only with direct or non-derivative culpability.[38]

Pereboom's response is somewhat ambiguous. He initially claims that such a response is dialectically unsatisfactory because it explicitly appeals to a 'leeway position in support of its verdict about Joe's responsibility.'[39] He claims that making this move risks failing to see the force of an important counterexample. However, Pereboom does also concede that the derivative responsibility move 'may get matters right.'[40] One way to respond to Pereboom's worry about dialectical status here is to point out that if we we can give an alternative explanation of the intuition we have that Joe is responsible in this case, i.e. the derivative move as Widerker contends, then the Frankfurt defender must show that this is not in fact the proper account of what's going on. Isn't that alternative explanation of the intuition exactly what the Frankfurt defender would want to rule out here? It is also a notably non *ad hoc* possibility that might explain these intuitions. It is not in any way dialectically inappropriate to raise this as a possible explanation. That said, could these competing explanations of the intuition not constitute a dialectical impasse here be-

---

[38] Widerker (2006: 173). Ginet (1996) anticipates this objection.

[39] Pereboom (2014: 19)

[40] Pereboom (2014: 19)

tween the leeway theorist and the Frankfurt defender? Well not if we have the resources to be able to tell whether our intuitions in buffer cases are tracking direct or derivative responsibility and we plausibly do have just those resources in so far as we can compare and contrast the structure of these cases with paradigmatic examples of derivative responsibility.

But how plausible is this response? Is *Tax Evasion 2* really an instance of derivative instead of direct moral responsibility? Pereboom cites an important disanalogy between the classic case of derivative responsibility with the drunk assault and Joe's position in *Tax Evasion 2.* In the case of the drunk assault outlined above, the person knows before they drink that given the likely outcome of drinking and socialising, they are most likely putting themselves in a situation where there won't be robust blameless alternate possibilities of action if they drink. By contrast, when Joe in *Tax Evasion 2* is not sufficiently attentive to moral reasons (this is the purported analogue of the drinking), he does not know that this lack of further attention will ensure the final decision he'll make will go a certain way. This disanalogy, if right, is surely key, for it seems that, during his inattentive deliberation, Joe wouldn't be knowingly doing something that will likely lead to blameworthy outcomes, as with the drunk assault case.

But is this disanalogy right? In reply it might be countered that Joe is doing exactly that. It could be argued that not being properly attentive to moral reasoning is just the sort of thing that can lead to trouble and,

what's more, that anyone capable of being morally responsible knows this full well. However, although not being attentive to moral reasoning puts you at a greater risk of acting in morally undesirable ways, it doesn't do so in the same way as drinking alcohol in the drunken assault case already described. There is at the very least a significant difference in probability here. With the drunken assault case, the chances are very high that you'll go on to act in a blameworthy way, that's kind of a done deal. But is this really true of Joe in *Tax Evasion 2*? Again, Joe is aware that at any moment he can refocus and entertain more moral considerations. Joe also still thinks it's completely up to him whether he goes on to cheat or not in this case. Lastly, I think that the moral attentiveness that Joe *does* have in the example, which is already (by stipulation) sufficient to meet the epistemic requirements on moral responsibility when he goes ahead and cheats on his taxes without intervention, means he is not violating any general conditions on deliberation as the drunk is clearly doing when they choose to drink. Even though it's possible that Joe could have been *more* attentive and Joe (like anyone) knows that the more attentive he is the *better* the chance of making the morally decent decision, it's still true that if Joe can believe his level is perfectly sufficient to make the morally decent decision then this is at odds with the paradigm cases under discussion. These disanalogies suggest to me that it's fair to conclude that the intuition we have that Joe is morally responsible is not best explained by the derivative model here. However, none of this is to say that upon closer inspection, Joe in *Tax Evasion 2*, although directly moral-

ly responsible, still doesn't have access to robust alternative possibilities. That is to where I now turn.

1.8 The timing criticism

Carl Ginet had also responded to the original *Tax Evasion* examples, in his review of Pereboom's *Living Without Free Will*.[41] The same line of criticism that Ginet develops there is applicable against the more recent *Tax Evasion 2*. Building on earlier work from his 1996 paper on Frankfurt cases, Ginet points out that there is a distinction to be made between (i) knowing that by choosing to attend to moral reasons one wouldn't know one was thereby avoiding responsibility for what might end up happening - or end up happening by some deadline, and (ii) knowing that by choosing to attend to moral reasons at time *t* one was avoiding responsibility for choosing to evade taxes at time *t*.

Ginet claims that if the second sense is at issue, then Joe's alternative (attending to moral reasons) *is* robust. This is because it is clear that by attending to moral reasons at *t* Joe would know that it was impossible that he was also choosing to evade taxes at the same time and hence he would be aware that by choosing to attend to reasons at time *t* he would be consequently avoiding responsibility for choosing to evade taxes at time *t*. This point can be run for any time *t* and for any buffer condition you might stipulate. It seems then that there is a robust alternate possibil-

[41] Ginet (2002)

ity present in *Tax Evasion 2*. On Ginet's approach, any old bit of deliberation at time *t* (whether or not the buffer of higher moral attentiveness is crossed) and where a decision to evade hasn't yet been made would thus constitute a robust alternative: the alternative *to refrain* at *t*. Similar points to Ginet's here were developed by David Palmer and Christopher Franklin.[42] Both argue that there are robust alternatives when time indexing is factored into the description of the case.

In summary then, although perhaps Pereboom is right about the disanalogy with cases of derivative responsibility, these arguments from Ginet, Palmer and Franklin undermine Pereboom and Hunt's claims that there are no robust alternatives to choosing to evade taxes in *Tax Evasion 2* and analogously for *Revenge*. So when Joe does in fact choose to evade taxes at some time *t*, he could have availed himself of these robust possibilities instead. If true, the timing criticism would undermine the claim that these buffer cases refute the principle of alternative possibilities.

Worries about the timing criticism and the scope of responsibility

Hunt and Shabo have defended the buffer strategy by taking issue with the timing criticism.[43] Their central worry is that the timing criticism forces us to deny that agents in Frankfurt cases who act on their own (i.e. when the counterfactual intervention doesn't occur) are responsible for their actions *simpliciter*, with them being instead merely respon-

[42] Franklin (2011), Palmer (2011)

[43] Hunt and Shabo (2013)

sible for *acting that way at t*. Hunt and Shabo contend that this is at odds with how we standardly understand responsibility for avoidable action. The move to restrict responsibility to times, they argue, is therefore *ad hoc* and purely in the service of preserving PAP. Moreover, when we instead rightly focus on responsibility *simpliciter*, they argue that the buffer examples still constitute counterexamples to the principle of alternate possibilities.

In what follows I will evaluate the arguments Hunt and Shabo put forward against the timing strategy and concede that we shouldn't relativise responsibility to times as they say. However, despite this, I go on to develop a new line of argument defending the basis of the move Ginet makes about robustness in buffer cases *without* having to relativise responsibility to times. Building on earlier work about whether agents in Frankfurt cases can be coherently said to *act* in the counterfactual case (as developed by Steward, Alvarez and Larvor) I develop a response to Hunt and Shabo which combines those insights about the counterfactual case with some new considerations about the epistemic requirements on robustness with respect to responsibility *simpliciter*. In short, Hunt and Shabo might be right about the importance of responsibility *simpliciter* but this won't ultimately help the buffer strategy as I will argue. It won't help because we can happily combine the idea that agents in buffer cases are responsible *simpliciter* for their actions with the claim that they nevertheless have robust alternatives to those actions. I further claim that the possibility of making the line of argument I develop has been overlooked

precisely because all sides in the timing criticism dialectic share an assumption (the very assumption that Alvarez, Steward and Larvor have given us good reasons to reject) that the agents in buffer cases *still act* in the counterfactual case. Once that assumption is dropped, I will argue that we have the means to successfully defend the principle of alternative possibilities against the buffer strategy.

Hunt and Shabo start by insightfully pointing out that it even if Ginet, Palmer and Franklin are correct about there being a robust possibility in the time restricted sense described above, it had better not *also* be the case that the agent in buffer cases is morally responsible for what they do in the actual sequence *simpliciter*. This is because then there will be some action which is both unavoidable for the agent *and* for which (in the actual sequence) the agent is morally responsible. But of course if that were true then we would still have a straightforward counterexample to the principle of alternative possibilities. Even if there *is* a robust alternative to being responsible *at t*, this won't be of any consequence if the agent is nevertheless responsible *simpliciter* for an action they couldn't avoid.

Hunt and Shabo make two main points to illustrate it is wrongheaded to restrict responsibility to times. The first involves paying attention to the same kind of action being performed in almost identical circumstances apart from at slightly different times:

> (Z1) Zeke has broken into Chad's office, where he hopes to surprise and kill Chad as part of his plan to eliminate a key witness in the case against him. When Chad enters the office alone a short while later, Zeke steps out from the shadows, gun in hand. Suppressing pangs of

> conscience, he stares into the other man's eyes for a full second before deciding to pull the trigger without further ado.[44]

They then consider variation Z2, where everything happens as in Z1 apart from the fact that the trigger is pulled two seconds later rather than one. In Z3 Zeke doesn't wait for the elevator and arrives at the office ten seconds earlier and shoots Chad straightaway. They contend that what all these cases clearly have in common is that Zeke is responsible for killing Chad:

> Our contention is that these further, temporally specific attributions of moral responsibility are of little account. Barring special assumptions, once it's been noted that Zeke is morally responsible for deciding to shoot Chad as part of his plan to eliminate a witness, noting that Zeke is also morally responsible for so deciding at t in Z1 (and for so deciding at t + 1 in Z2) adds little if anything of consequence….[45]

It might be said in response to Hunt and Shabo here that this isn't exactly knockdown as Palmer and Franklin could presumably just repeat that what we are *strictly* responsible for is the time indexed event. That said, the burden is indeed on Palmer and Franklin to establish this as opposed to the more plausible attributions of responsibility *simpliciter* as Hunt and Shabo rightly contend is the more natural account of what's going on here. The point is that in a standard case of avoidable action the presumption is that the agent is responsible *simpliciter*. Apart from the fact that the move to restrict responsibility to times allows the defender of

---

[44] Hunt and Shabo (2013: 606)

[45] Hunt and Shabo (2013: 607)

PAP to block the Frankfurt defender here, Palmer and Franklin have not argued it is independently well motivated.

The second point Hunt and Shabo make is much more telling I think. They consider the spatial analogues of how responsibility would look if it was always indexed to the particular *place* you acted:

> The prima facie moral triviality of the timing of Zeke's decision is of a piece with the moral triviality of other properties of Zeke's decision, making it all the more incumbent on Palmer to explain why the exact time at which an action is performed should play the unique role he assigns to it. In Z1, for example, Zeke not only decides at t to pull the trigger; he also so decides from behind the desk, standing ten feet from Chad, with a bead of sweat running down the bridge of his nose, as a car backfires on the street, while in Z2 he not only decides at t + 1 to pull the trigger; he also decides near the filing cabinet, standing nine feet from Chad, with a bead of sweat hanging from the tip of his nose, as a second backfire is heard from the street. Just as Zeke is morally responsible in Z1 for deciding at a certain point in time (t) to shoot Chad, so he is morally responsible for deciding at a certain point in space (behind the desk, ten feet from Chad, and so on) to shoot him. But it would be absurd to suppose that in ordinary cases of avoidable action the location of the decision is an essential part of what the agent is morally responsible for, and that in a Frankfurt version of Z1 Zeke wouldn't be responsible for deciding to shoot Chad simpliciter but only for so deciding at the exact location where the decision is made.[46]

This seems right. There are in fact many properties that seem irrelevant to the responsibility ascription and it is hard to see why the precise time at which a certain type of act occurs should be treated differently to these other properties (such as spatial location) that are intuitively of no consequence. These are standard cases of avoidable action and it would be *ad hoc* to insist things were different for unavoidable action, as Hunt and

[46] Hunt and Shabo (2013: 608)

Shabo say. This leaves the leeway defender in a tight spot as if responsibility *simpliciter* is what the agent bears then we still have a counterexample to PAP for the reason described above. However, I contend this can be remedied by paying attention to the status of the counterfactual case and the intervention that takes place there as well as the proper epistemic requirements on robustness.

1.9 A new response to the timing criticism dialectic: Defending the principle of alternate possibilities against Hunt and Shabo

The above section concludes with the thought that what goes for avoidable action should go for unavoidable action, i.e. it's responsibility *simpliciter* in both cases. Although I think Hunt and Shabo are right about responsibility *simpliciter* being the responsibility attribution that matters, I don't think that there are instances of unavoidable action in Frankfurt cases. The problems of the status of the counterfactual case in Frankfurt examples where manipulation occurs have already been discussed following the examination of Alvarez, Steward and Larvor's points earlier in this chapter. Manipulated agents don't *act* in Frankfurt cases. What has not been noted in the literature to date is that the upshot of this point for the dialectic of the timing criticism means there is never anything that the agent is responsible for that wasn't avoidable. If this is correct, then crucially, the leeway incompatibilist does not need to relativise to times in order to hang on to the avoidability requirement that is central to the

leeway position. Just as well given Hunt and Shabo are right about the problems with restricting responsibility to times. The problem, as it seems to me, is that people on both sides of the debate about buffer examples have been (tacitly or otherwise) accepting that although the agent isn't responsible when they are manipulated, they still act in that scenario. It is very interesting to see how many commentators take this for granted in these cases including, crucially for my argument here, those trying to *defend* the principle of alternative possibilities in the timing dialectic. Ginet himself, in characterising the general schema of Pereboom's original *Tax Evasion* case says, "The Intervenor, in order to ensure that Jones did B at t1, set up a backup mechanism such that, if C had occurred in t0 − t1, the mechanism *would have caused Jones' doing B at t1.*"[47] Having set out the schema *he* thinks *Tax Evasion* instantiates, and not just what he's reporting Pereboom claims it instantiates, Ginet continues: "The PAP defender must further claim that, although Jones was morally responsible for his doing *B* at t_1 (having had a robust alternative), he is not responsible for the less specific fact, *for which he had no alternative, that he did B by* t_2."[48] In Palmer we find, "In my view, while Joe can be responsible for deciding to evade taxes at t1, he cannot be responsible for the more general fact that he decided to evade taxes simpliciter. *This is because, due to the neuroscientist's presence, this more general fact is not something that Joe*

[47] Ginet (2002: 307) my italics.

[48] Ginet (2002: 308) my italics

could have avoided."[49] In Franklin's 2011 discussion he says, "… if Jones had considered not killing Smith at t_3, *Black would have intervened and forced Jones to decide to kill Smith at some later time*…"[50] It's no different with Hunt and Shabo themselves. Indeed, a necessary part of Hunt and Shabo's criticism of the timing move is based on the idea that it's responsibility *simpliciter* that counts *and agents in buffer cases are responsible simpliciter for actions they couldn't avoid given the presence of manipulation mechanisms and interventions*. Their argument against the timing move relies on the assumption we should reject.

There is thus a consensus between the opposing parties in the timing debate on the coherence of unavoidable action in the sense at issue when manipulation and intervention occur in Frankfurt cases. Whether in the actual sequence (where the agent acts responsibly on their own), or the counterfactual case (where although all agree the agent doesn't act responsibly *they think he still acts*) all agree there is an action performed by an agent where that action is unavoidable for that agent given the structure of the situation. I claim this consensus is mistaken and should be rejected for the reasons already put forward by Steward, Alvarez and Larvor. Furthermore, while I reject the claim that the agent in a Frankfurt case acts in the counterfactual scenario when they are manipulated, I can still employ Ginet's observation about the robustness of refraining *at any particular time t* while maintaining the agent is nevertheless responsible

[49] Palmer (2011: 269) my italics

[50] Franklin (2011: 197) my italics

simpliciter for their decision at that time and not merely responsible *for their decision at t.* Responsibility *simpliciter* and avoidability are both maintained in the intuitive way here. I can agree with Hunt and Shabo on responsibility *simpliciter* and make use of Ginet's insight about the *awareness* of not being responsible for something *you're not then doing* to preserve the intuitive leeway incompatibilist picture here.

A defence of the combination of responsibility s*impliciter* and robust alternatives in *Revenge* and *Tax Evasion 2*

It might be said that even if the agent in a buffer case doesn't act in the counterfactual scenario and that as a result there is no action the agent is both morally responsible for and which is also unavoidable, this doesn't necessarily mean the alternative is robust in the way required for responsibility *simpliciter* as opposed to mere responsibility for an act at a specific time. I think it's clear enough that if *responsibility at a time* was what was at issue then the observation Ginet makes in his original statement of the timing criticism would decisively mean that the alternative is robust. That is, it's clear that when I refrain/deliberate at some time *t* or cross the buffer at some time *t*, *I will be aware* that at *t* I'm doing something such that I will not be responsible for *avoiding tax at t* etc. The crucial question now is whether this awareness of the status of the situation is sufficient to make the alternative robust with respect of the attribution of responsibility *simpliciter.*

In response to the above worry, it seems perfectly coherent to think that at any point in time when I'm *not* making a decision for which I would be morally responsible (were I to make it) and I know this (in line with Ginet's original point about the agent's awareness) I can coherently be committed to the thought that I would have been responsible *simpliciter* for that decision and action had I made that decision right then instead. When I act in such a way that I'm morally responsible I know that I didn't have to act that way *then or ever.* I think this about every moment of time continuing into the future *when it's still up for grabs what I will eventually do as well as when I have committed myself to a course of action*. Even in the latter case I standardly don't think my commitments are unavoidable. One of the worries in the literature was that I would need to know that (in order for it to be robust) the alternative would rule out responsibility for what would *end up* happening and that (given that deliberating and merely crossing the buffer do not commit me epistemically either way), the alternatives in *Revenge* and *Tax Evasion 2* are not robust. I think this assumption should be rejected. Although I may often commit myself to acting a certain way over time or in the future, an *alternative* to that course of action does not *itself* need any temporal commitments built into its epistemic component in order to be robust.

Why would anyone think that alternatives (in order to be robust and exempting) *had* to commit us to not do whatever bad action we were contemplating at times continuing into the future? Of course sometimes we could have chosen an alternative path that included a short, medium or

long term commitment to do something else but the point is these variable commitments of the alternatives are not important with respect to the question of whether they are robust. People can always change their minds and frequently do change their minds. New or previously neglected reasons can at any point involuntarily weigh in on us and cause us to reconsider and change our minds. This can happen seconds after a decision has been made, although of course we often do stick by the courses of action decided upon. Because of this it would be wrong to require of an alternative anything other than that taking it absolves us (right then) of responsibility *simpliciter*. We may seconds later change our minds and then incur that responsibility for which we avoided only seconds earlier. All of this, as I said in the above paragraphs is consistent with the fact that when we do incur that responsibility, it's responsibility *simpliciter* and not responsibility indexed to times. Hunt and Shabo focus on this very point when they discuss how we should understand the epistemic component of robustness. They consider the two possible readings of Pereboom's robustness condition that Franklin outlines. The first being the narrow reading that the agent only needs to know that by taking the alternative at some time t they will not be responsible for the action at that time t and the stronger requirement that by taking the alternative the agent must understand that they will avoid responsibility for that action at all subsequent times into the future. Hunt and Shabo concede that the stronger reading should be rejected:

> On the second, stronger reading (ibid.), an agent must understand that, by realizing the alternative possibility, she will avoid being responsible for her action *then and at all later times*. It seems clear that Jones doesn't meet this requirement; for all he knows, he might well decide to kill Smith immediately after considering (at t) not killing Smith.
>
> We agree with Franklin that EC seems implausibly strong on the second reading. Whatever else, having a robust alternative possibility doesn't require an agent to understand that, if she realizes that possibility, she will *never* be morally responsible for a decision like the one she actually makes.[51]

Although they don't elaborate on why this is, presumably it's because the future is uncertain and people change their minds. But the crucial observation here is that this point is true of *all* future times, not just the more distant times. Because of this it would be wrong to build any temporal commitments into the robustness requirement. This is true even when people have in fact committed to a course of action over the long term. The point is that a robust alternative to them doing *that* need be nothing over and above them simply not doing that then, with them instead simply continuing to deliberate or whatever.

Further evidence that there need be no temporal commitment about future actions or omissions built into the epistemic component of the alternative is to be found by re-examining Pereboom's own robustness criterion: Robustness (B):

> **Robustness (B)**: For an agent to have a robust alternative to her immoral action A, that is, an alternative relevant per se to explaining why she is blameworthy for performing A, it must be that

[51] Hunt and Shabo (2013: 613)

> (i) she instead could have voluntarily acted or refrained from acting as a result of which she would be blameless, and
> (ii) for at least one such exempting acting or refraining, she was cognitively sensitive to the fact that she could so voluntarily act or refrain, and to the fact that if she voluntarily so acted or refrained she would then be, or would likely be, blameless.[52]

It requires no argument to see that there is nothing whatsoever in Pereboom's Robustness (B) criterion that is at odds with what I have been claiming above. Robustness (B) simply does not entail any temporal commitments should be included in the epistemic component of the alternative in order for that alternative to be robust and in consequence exempting in the intuitive way. All that is required, to summarise here, is that there needs to be something you could have done instead that (i) would have rendered you blameless and (ii) you were aware that that alternative would have likely rendered you blameless. In conclusion, these points can both be satisfied at every moment as an agent moves through time deliberating what to do and not to do while requiring no temporal commitment other than a commitment not to have done the action for which you are responsible *simpliciter* when you in fact do it.

Given these clarifications, it is useful in conclusion to return to Hunt and Shabo's response to Ginet, Palmer and Franklin in order to illustrate how the leeway incompatibilist should now respond armed with these considerations. Hunt and Shabo (working on the correct assumption)

---

[52] Pereboom (2014: 13)

that it's responsibility *simpliciter* that counts and (the false assumption) that agents in Frankfurt cases still act when manipulated say the following:

> On the other hand, if Jones is morally responsible for deciding to kill Smith, it's hard to see what turns on his understanding that he can avoid responsibility for so deciding at t. If he is indeed morally responsible for so deciding simpliciter, then considering not killing Smith in Black's absence before deciding on his own to kill Smith wouldn't absolve him of responsibility for so deciding. Why think, then, that what grounds or explains his responsibility for deciding to kill Smith in Black's presence is his understanding that he can avoid responsibility for so deciding just then? Understanding that he can avoid responsibility for so deciding then doesn't make his decision avoidable in any interesting sense.[53]

A little later we find:

> More generally, if a PAP-defender claims to identify a robust alternative possibility in a Frankfurt case, we believe that it's fair to ask her whether, in Black's absence, exploiting that alternative possibility before adopting the course she actually does adopt would eliminate (or mitigate) her blameworthiness. If the answer is ''no,'' it is especially incumbent on her to explain why that unutilized ''possibility space'' should make the crucial difference to the agent's responsibility.[54]

And finally:

> Perhaps there is a good answer to this question in the case of decisions. Perhaps, that is, there is room to explain why being able to decide against killing Smith should be considered a robust alternative, even if exploiting this possibility in Black's absence before deciding to kill Smith after all would not lessen Jones's blameworthiness. Be this as it may, we believe that the question is a good one, and that there is a presumption against any view that cannot give it an affirmative answer. If exploiting the residual leeway in Black's absence would not

[53] Hunt and Shabo (2013: 614)

[54] Hunt and Shabo (2013: 619)

> lessen Jones's blameworthiness, why think that his actual responsibility in Black's presence hinges on this unexploited leeway?[55]

In response, the agent in a Frankfurt case doesn't have to act that way at that particular time or *any* future time so when Hunt and Shabo say "… Understanding that he can avoid responsibility for so deciding then doesn't make his decision avoidable in any interesting sense", they are simply assuming that the action is in the end unavoidable. But this is precisely what I reject for the reasons already given regarding whether the agent acts in the counterfactual case. From the second quotation, they contend that it is a fair question to ask whether exploiting the earlier alternative possibility would diminish responsibility when the agent ends up acting a certain way later in Black's absence. If the answer is no they contend it seems hard to see why the same earlier alternative possibility should diminish responsibility for what you end up doing when Black is present in a Frankfurt case. Although the answer to the question they ask is indeed 'no' in the first case, this is unproblematic because when Black is present there is never an action of the agent for which they are responsible that was unavoidable. If they acted 'on their own' with Black present then they are responsible but they didn't have to act that way. If Black manipulates we no longer have a candidate for an *action of that agent*. Another way of putting this would be to say that although the answer to their question is 'no' when Black is absent, when Black is present the answer is still unproblematically 'no' when the agent acts without

[55] Hunt and Shabo (2013: 620)

Black getting involved or, the question is simply inappropriate if Black intervenes as then the agent doesn't end up doing anything. Agents are responsible when they freely and avoidably choose to act as they do. They don't act when manipulated by Black so as to causally guarantee outcomes. Finally, and along the same lines, responding to the question in the third quotation above, the 'unexploited leeway' is significant because it is in virtue of that unexploited leeway that the agent is responsible in *both* cases where Black is absent *and* present, i.e. in the cases where the agent acts *without* manipulation occurring. So I can happily agree that the leeway isn't treated asymmetrically depending on Black's presence as Hunt and Shabo rightly contend it shouldn't be here.

In summary, I conclude I have shown that a significant breakthrough is possible in the timing criticism dialectic for the leeway incompatibilist against the Frankfurt defender. The timing criticism and the issue of whether we should index responsibility to times is one of the most important ongoing debates relevant to the whether the buffer strategy, which is the most sophisticated Frankfurt-style strategy to date, can be made to work. I have shown that because of a consensus in the literature between certain leeway advocates *and* Frankfurt defenders regarding the status of the manipulated agents in buffer cases (their status *as agents* in the counterfactual case) a way of defending the leeway view has been overlooked. I have shown that, armed with the arguments from Alvarez, Steward and Larvor concerning the status of the counterfactual case, as well as the appropriately stated epistemic construal of the robustness cri-

terion, the leeway theorist can make progress here and defend the principle of alternative possibilities in an intuitively pleasing way. The leeway incompatibilist can happily agree with Hunt and Shabo that agents are responsible *simpliciter* for what they do but she can at the same time defend the claim that those agents still have robust alternative possibilities for the actions for which they do incur responsibility. In conclusion the classic buffer cases of *Revenge* and *Tax Evasion 2* pose no threat to the principle of alternative possibilities. So far so good, I now turn to the very latest buffer case that Pereboom as developed in light of the timing criticism.

1.10 *Tax Cut*

Pereboom has recently constructed a new buffer case *Tax Cut*, which aims to avoid Ginet's timing criticism. *Tax Cut* runs as follows:

> ***Tax Cut***: Jones can vote for or against a modest tax cut for those in his high-income group by pushing either the 'yes' or the 'no' button in the voting booth. Once he has entered the voting booth, he has exactly two minutes to vote, and a downward-to-zero ticking timer is prominently displayed. If he does not vote, he will have to pay a fine, substantial enough so that in his situation he is committed with certainty to voting (either for or against), and this is underlain by the fact that the prospect of the fine, together with background conditions, causally determines him to vote. Jones has concluded that voting for the tax cut is barely on balance morally wrong, since he believes it would not stimulate the economy appreciably, while adding wealth to the already wealthy without helping the less well off, despite how it has been advertised. He is receptive and reactive to these general sorts of moral reasons: he would vote against a substantially larger tax cut for his income group on account of reasons of this sort, and has actually done so in the past. He spends some time in the voting booth rehearsing the relevant moral and self-interested reasons. But

> what would be required for him to decide to vote against the tax cut is for him to vividly imagine that his boss would find out, whereupon due to her political leanings she would punish him by not promoting him to a better position. In this situation it is causally necessary for his not deciding to vote for the tax cut, and to vote against it instead, that he vividly imagine her finding out and not being promoted, which can occur to him involuntarily or else voluntarily by his libertarian free will. Jones understands that imagining the punishment scenario will put him in a motivational position to vote against. But so imagining is not causally sufficient for him to decide to vote against the tax cut, for even then he could still, by his libertarian free will, either decide to vote for or against (without the intervener's device in place). However, a neuroscientist has, unbeknownst to him, implanted a device in his brain, which, were it to sense his vividly imagining the punishment scenario, would stimulate his brain so as to causally determine the decision to vote for the tax cut. Jones's imagination is not exercised in this way, and he decides to vote in favor while the device remains idle.[56]

The aim of *Tax Cut* is to eliminate the possibility of making Ginet's timing move right at the end of the 2 minute period. If successful this would mean that Jones wouldn't be able to refrain at the end of the two minute period and hence PAP would be challenged as Jones is still intuitively morally responsible when he votes for the tax cut at the last possible moment. Up until the last minute possible, it will of course be open to the PAP defender to make the same move as before, i.e. following my suggestions in the section above, there are robust possibilities with or without crossing the buffer by just deliberating at any time t before the deadline. As I said in the previous section, these alternatives are robust as they meet the epistemic requirements and furthermore, responsibility (when it is incurred) is responsibility *simpliciter*. What about at the last

[56] Pereboom (2014: 23)

point at which it is possible to make a decision before the timer runs out? Jones would have known that he had to make a decision right then and given the fact that the decision to not vote is not causally possible (by stipulation), Jones will not have a robust alternative to his choice to vote for the tax cut for which he is intuitively morally responsible.

Buffer crossing and intervention

One worry that appears to be relevant here concerns whether there is a time lag or not between the point in time when the buffer is crossed and the point when (once crossed) the ability to do otherwise becomes a live possibility and hence a robust possibility. If there was a time lag of any duration then presumably we could just stipulate into the example that the intervention process would be able to intervene and manipulate the agent during that time lag and hence before the agent had access to the robust alternative, in line with Pereboom's approach here. But what about if, as soon as the buffer was crossed, the robust alternative simultaneously became an option? That doesn't seem to be incoherent and is perhaps even what we might expect here. If this was the case, then given the fact that the buffer is operating as a prior sign we know it must take some time (however short an interval) for the detection process to trigger the intervention process, we might have a window at which the agent could have access to a robust alternative, however short a window that might be, and that is going to be a problem for *Tax Evasion 2* and *Tax Cut*. In *Tax Evasion 2*, it would mean that Joe could have chosen not to evade

as soon as he crossed the buffer for a short period of time before the inevitable intervention. What about in *Tax Cut*? *Tax Cut* is harder because Pereboom's description of the case makes it seem like Jones will somehow (if he has left the decision right to the last possible moment) not suddenly cross the buffer *because* that would mean letting the time window slip by instead of choosing to vote for the tax cut. However, if I'm right that the possibility of choosing to vote against can become a live possibility simultaneously with his crossing the buffer (imagining the boss in this case), then even if the buffer was crossed at the last instant, *it would then be possible for Jones to vote against at the last instant*. Why should we think (given Jones' understanding of his situation), that he would think he had to vote for the cut at the last moment any more than he had to vote against it (ignore for the moment whether it was possible for him to make the choice to vote against)? This last point taken together with the suggestion of the robust alternative becoming a live option simultaneous with the buffer crossing would mean that even in Tax Cut Jones might well have a robust alternative *up to and including* the last moment to choose to vote against.

All of the above seems coherent as a possibility but is it plausible? Perhaps it might be said that there has to be a period of time after the buffer has been crossed in order for deliberation to 'digest' the new considerations, for the agent to mull over the options again in light of the new found sensitivity to moral reasons. I want to say two things in response to this possibility. Firstly, I don't see why one would insist on this.

Specifically, why doesn't crossing the buffer immediately put the agents in just that state of heightened sensitivity necessary to make an properly informed decision? Why insist on a time lag? Secondly, suppose that it were conceded here that there would need to be a small time lag between the buffer being crossed and the time at which the robust alternative became a live possibility so to speak. Even so, it should be noted that we are now working in a prior sign context again. Because of this, we might well ask why we should expect the period of time that it takes for the robust possibility to become truly available after the buffer is crossed to be longer than the *necessary* time lag before the intervention process is triggered after it's been detected that the buffers have been crossed. It looks like we are at a point in the dialectic where it is merely stipulation one way or another that could settle such a question.

However, perhaps imagining the buffer cases in a possible world where the time lags come out the way Pereboom needs them to is just as legitimate as I found the stipulation of the buffers in the first place. I argued above that even if our ordinary concept of what libertarian free will involves is in tension with the stipulation of the buffer in general, Pereboom can nevertheless legitimately ask us to imagine the closest possible world where the buffers did work (by stipulation) in the way he wanted. Now, analogously, he can also ask us to imagine a world where (again by stipulation) we needed a period of time to deliberate post crossing the buffer before we could access robust possibilities. Even if this world differed from how some libertarian possible worlds are structured, Pere-

boom can concede this and nevertheless ask us to consider the world where the set up in his example works. There is nothing conceptually problematic about this stipulation at some possible world, even if there is given our actual conception of libertarian freedom and what that concept rules in and out in the actual world. As I said above, how the actual concept of libertarian free will works here (if indeed there is a uniform, stable concept) is not necessarily a guide to whether robust alternates are *conceptually necessary* for moral responsibility. So far then, for all the buffer cases examined, the buffer stipulations and associated time lag stipulations necessary to make the examples work are not problematic in themselves. We must look to other features of the examples.

1.11 A Dilemma for *Tax Cut*

One of the crucial aspects of the buffer cases to get clear on is the specification of what the agent who is the subject of these cases knows regarding their situation. There are two aspects that need to be clarified. Firstly, do the agents know about the counterfactual intervention should they deliberate/think a certain way (i.e. do they know they're in a Frankfurt case) and secondly, do they know about the buffer (irrespective of knowing about the counterfactual intervention)? The answer to the first question is uncontroversially 'no'. It is part of the structure of Frankfurt cases that the intervention mechanism is hidden and doesn't play any causal role in the actual sequence. The second issue is harder though and it isn't immediately obvious from Pereboom's description of the cases what is

being stipulated. For example, in the description of *Tax Cut*, Jones understands that imagining the punishment scenario will put him in a motivational position to vote against the cut. Does this mean that Jones merely knows he'll definitely be able to vote against then or that (more strongly) it is a necessary condition of him being able to vote against that he imagines his boss and the punishment? It isn't specified but we can consider how *Tax Cut* looks in both the situation where Jones knows about the buffer and where he doesn't.

In the situation where Jones knows there is a buffer (functioning as a necessary condition), i.e. where he knew that he had to consider his boss's reaction if he was to be able to vote against; if he leaves the decision until the last possible moment and has not considered his boss's reaction by then he has knowingly put himself in a position where he is causally determined to vote for the cut. However, in this case the situation looks like a classic case of derivative responsibility. Jones omitted to do something at an earlier time that he knew he had to do (and could have done) if he was to avoid the bad outcome later. When we build knowledge of the buffer into the case and the buffer isn't crossed, we have a situation which does look paradigmatically like the drunk driver or inevitably violent drunk man in Pereboom's classic cases of derivative responsibility. But if this is the case then *Tax Cut* fails to provide a counterexample as derivative responsibility is now traced back to an earlier situation where there was a robust alternative possibility. The fact that (referring to that earlier decision not to cross the buffer to which de-

rivative responsibility is now arguably traced) buffer crossing would have been consistent (without the immediate intervention) with still being able to vote for the cut is not a reason to think buffer crossing is not robust here. As I have already made clear in the section on the timing dialectic., robust alternatives do not *need* to epistemically commit the agent to avoiding an alternative at times into the future (though they may also do this).

So at any point during the ticking timer's duration, in the scenario where Jones knows about the buffer being a necessary condition for voting against there is a robust alternative up to the last possible moment he can make a choice (i.e. his *not* choosing to vote in favour then) and at the last possible moment (if the buffer has not been crossed) Jones is causally determined to vote in favour but we get what looks like a paradigm case of derivative responsibility that traces back to earlier omissions (the omission to cross the buffer) for which there were robust alternatives available. Either way, there is no situation (when Jones knows about the buffer) where he is *directly* morally responsible for some decision for which he didn't have a robust alternate possibility.

On the other hand, if Jones doesn't know about the buffer, presumably he believes (falsely) that he can make the decision to vote against at any point up to *and including* the last possible moment regardless of whether he has considered his boss's response or not then. Before the last possible moment there is a robust possibility here (just as in the scenario where he knows about the buffer,) namely Jones not choosing to vote in favour at

that particular moment (as per the timing criticism). As for the deadline, if we take Pereboom's stipulation seriously that it is causally determined that Jones votes by the deadline, then given he can't vote against because he hasn't crossed the buffer, he will be causally determined to vote for the cut. But then the trouble is that Jones will have unwittingly (and without prior negligence) put himself in a position where he is causally determined to do something bad, for which many will have the intuition that he isn't responsible. If Jones didn't know that he needed to cross the buffer in order to have the ability to vote against, this suggests that Jones might have acted differently had he known these things.[57] We can imagine the same scenario as a 'torn decision' or, figuratively as a 'mental coin toss' where Jones is determined to decide by the deadline and (not knowing about the buffer) doesn't have a preference either way. Once again, if we are to take Pereboom's stipulation that it is causally impossible for Jones to decide to vote against if he hasn't crossed the buffer seriously, then in the 'mental coin toss scenario' where the buffer has not been crossed we will have a situation where the decision made by Jones will always come out one way, namely to vote in favour of the cut. Once again, it seems that this is not something for which Jones is intuitively responsible for the same reason as mentioned above. We can imagine what Jones would subsequently say if told about the buffer. He might well say "that so called torn decision was rigged to come out that way

[57] I am grateful to Carl Ginet and Nadine Elzein for suggesting to me in correspondence that the causal process resulting in Jones' voting for the cut at the deadline needn't be something for which he is intuitively responsible.

given the set up"…. or "I'll stand by my freely made torn decisions but not decisions rigged like this" etc. In summary, on this horn of the dilemma, where Jones doesn't know about the buffer, there is no scenario where he is directly morally responsible for something for which he had no robust alternate possibilities. Before the deadline, if Jones votes in favour of the cut there are robust alternatives (as per the timing criticism). At the deadline it would appear that Jones is not intuitively morally responsible for what transpires after all given it will be causally guaranteed to happen.

1.12 Further problems with the structure of Tax Cut given the conditions on decision and action

In the section above, I criticised *Tax Cut* in the form of a dilemma and concluded that there is no version of the scenario playing out where Jones makes a decision to vote for the tax cut for which he is directly morally responsible and for which he doesn't have a robust alternative possibility. This is I believe the correct analysis for a leeway incompatibilist who believes that Jones *still acts* in all these scenarios. However, it should be noted that this analysis is at odds with my earlier claims about *Revenge* and *Tax Evasion 2*, where, given the analogous application of the arguments about classic Frankfurt cases by Steward, Alvarez and Larvor, I claimed the agent doesn't act in the counterfactual case when they are manipulated and intervention occurs. This raises the question whether those same arguments can also be applied here. And it seems they can.

So we now need to examine whether Jones acts in *Tax Cut* when the causal set up ensures the outcome (given the fact the buffer hasn't been crossed and the necessity of acting by the deadline). Before the deadline, if Jones chooses to vote in favour on his own in the normal way then he does of course act, is of course responsible but he has a robust alternative. If he crosses the buffer however, then just as before what happens next (given the intervention and causal determination of the subsequent state of affairs by the mechanism), we do not have a viable candidate for an *action of Jones*. This is for exactly the same reasons as given by Alvarez, Steward and Larvor. So this covers all the possibilities up until the decision deadline. At the decision deadline itself, given that the buffer will *not* have been crossed (if it had we would already have had an earlier case of 'non action' given the intervention) we have a situation where the lack of buffer crossing and the deadline causally necessitate the outcome irrespective of the state of Jones' deliberations and emotional states by then. I therefore contend that this process should also not qualify as a decision of Jones for just the same reasons, i.e. the so called 'decision' that comes about is not related to the earlier deliberative states in the right kind of way to qualify as an act of Jones. It is a little harder to see this at first, as the 'decision guarantee' mechanism in *Tax Cut* doesn't immediately appear to be of the same invasive kind as in the classic counterfactual case where Black swoops in and causally determines the outcome. However, if we take Pereboom's stipulations about the buffer and the necessity of acting by the deadline seriously, then the case (at the deadline)

looks to share all the properties that ruled out the agent acting in the classic counterfactual intervention scenarios. I conclude that *Tax Cut* doesn't provide a counterexample to PAP.

1.13 The 'What-should-he-have-done' defence and Otsuka's 'avoidability of blame' argument

So far in the discussion of the various Frankfurt cases I have been evaluating, I have only criticised the internal structure of the examples. That is, I have been considering whether, on closer inspection, the cases are coherent, or whether they have robust alternative possibilities in them after all. It's fair to say that the majority of the literature has proceeded in the same way. Commentators have got their hands dirty in the metaphysics of the examples as with Alvarez, Steward and Larvor's analysis of the counterfactual case, or argued that you can't rule out the alternative possibilities, as with the dilemma defence developed by Kane, Ginet and Widerker. There is however a distinct strand in the literature which takes a very different approach and argues directly against the purported conclusion of a (were it possible) coherent Frankfurt style case *without* challenging the details and set up of the cases, or at least without directly doing so. David Widerker and Michael Otsuka have both formulated arguments to this effect. These responses therefore directly challenge our having the intuition of responsibility about such cases, were they possible. Widerker has developed the 'What-should-he-have-done defence' or

the W-defence for short. He asks us to consider the following case, where the breaking of a promise was the salient action:

> Let me grant, for the sake of discussion, that in the IRR-situation under consideration, Jones acted freely in the sense that what he did he did for reasons of his own without being causally determined or coerced to so act. Still, since you, Frankfurt, wish to hold him blameworthy for his decision to break his promise, tell me *what, in your opinion, should he have done instead?* Now, you cannot claim that he should not have decided to break the promise, since this was something that was not in Jones's power to do. Hence I do not see how you can hold Jones blameworthy for his decision to break the promise.[58]

Widerker considers some possible objections to the W-defence and responds to them. Firstly, someone might argue that it's unfair to absolve the agent from Blame in a Frankfurt case because we can't answer the question, 'what should they have done instead?' They might suggest this because *for all the agent knew, they thought they could have done otherwise*. Widerker rightly notes that this response misses the point. If the only way to be absolved of responsibility was to have an excuse then this objection might be right. However, the defender of alternative possibilities may argue that a lack of alternatives also undermines responsibility. Given that alternate possibilities are taken to be necessary for responsibility, their absence would undermine responsibility regardless of whether or not the agent in question had an excuse. Widerker goes on to suggest that perhaps what the critic might mean is that the agent is still blameworthy because he *should have done what he believed was the right thing to do*. But if this is what is meant then Widerker can simply reply with the

---

[58] Widerker (2003: 63)

W-defence once again, namely that it is surely unreasonable to expect that an agent should do something they are not able to do.

Secondly, Widerker suggests that someone might try to counter the W-defence by taking the following line. Given that in a Frankfurt case what Jones' decision reveals about him and his character is exactly the same as it would be if the Frankfurt apparatus was removed (and thus where Jones has alternatives), the situations should be regarded as morally equivalent and we should consequently blame Jones in the Frankfurt case as we would in the normal scenario. In fact, this is a line that is strongly suggested by what Frankfurt himself says in the original 1969 paper. After introducing the most developed version of his classic case (Jones 4), Frankfurt says, "In that case, it seems clear, Jones4 will bear precisely the same moral responsibility for what he does as he would have borne if Black had not been ready to take steps to ensure that he did it. It would be quite unreasonable to excuse Jones4 for his action, or to withhold the praise to which it would normally entitle him, on the basis of the fact that he could not have done otherwise. This fact played no role at all in leading him to act as he did. He would have acted the same even if it had not been a fact. Indeed, everything happened just as it would have happened without Black's presence in the situation and without his readiness to intrude into it."[59] In reply, Widerker rightly points out that the two situations are *not* morally equivalent to someone who believes alternative possibilities are a necessary condition on moral

[59] Frankfurt (1969: 836)

responsibility. The asymmetry is precisely what the W-defence aims to draw attention two. While it would be reasonable to expect Jones to act otherwise in the normal situation, it looks unreasonable in the Frankfurt scenario. Widerker does concede though that saying this is consistent with us being able to assess Jones conduct negatively from a moral point of view: there are other forms of moral assessment apart from moral responsibility, or 'basic desert moral responsibility', as Pereboom would call it.

The reasoning in the W-defence is persuasive because, as I said at the start of this chapter, we move very naturally from the thought that someone is to blame to the thought that they should have done something other than they did in fact do. Coupled with the 'ought implies can' premise, we can then move back through these entailments by *modus tollens*, firstly from the fact that you couldn't have done otherwise to the negation of 'you should have done something else instead' and then in turn from the negation of that to the negation of the claim that the agent is blameworthy. The Frankfurt defender agrees that the agent cannot do otherwise but they also hold the agent blameworthy so they must challenge the *modus tollens* somewhere. Although highly counterintuitive, some Frankfurt case defenders here are prepared to bite the bullet and deny 'ought implies can'. I won't discuss these replies to Widerker here. I can happily agree with Widerker given I've argued that there are in fact robust alternative possibilities in the Frankfurt style examples. We should expect this perhaps given the strength of the W-defence. These

two lines of argument (the defence of the principle of alternate possibilities on the one hand and the W-defence on the other) are mutually supporting here. All well and good for the leeway incompatibilist then.

Michael Otsuka independently develops a response along the same lines: "I argue that blaming someone for what she has done is warranted only if she could have behaved less badly and that if she could have behaved less badly, then she could have behaved in a manner for which she would have been entirely blameless."[60] After a section explaining away intuitions that appear to support the view that we do hold people accountable and blame them in situations that look to undermine this principle, he says the following:

> Take any imagined pair of individuals who have behaved badly (e.g., who have maliciously injured another) and hold everything constant except for the fact that the one could have behaved less badly, and knew that she could have, whereas the other could not have behaved less badly. The fact that the one person behaved as badly as she did even though she knew that she didn't have to provides sufficient grounds for indignation in her case that are lacking in the second case. Moreover, there are no other grounds that are sufficient for indignation in this second case. Such grounds are lacking no matter how malevolent or otherwise vicious this person might have been.[61]

To my mind, this exercise brings into sharper focus the point that it seems groundless to blame the latter agent *once we know they couldn't have behaved less badly*. Proper reflection on the fact that an agent simply could not have behaved less badly raises a flag that would appear to be a prob-

---

[60] Otsuka (1998: 694)

[61] Otsuka (1998: 696)

lem for a Frankfurt case *even if it could be successfully described*, that is, even if such cases are structurally possible. Careful reflection and focus on the fact of unavoidability challenges the initial intuition of responsibility we tend to have in a Frankfurt case. This is a problem that has been largely ignored in the literature. Most work in this area proceeds on the assumption that the intuitions of responsibility in Frankfurt cases are secure and stable. To my mind, Widerker and Otsuka have shown that even if we granted the structural success of Frankfurt cases, the Frankfurt example constructor would still have to defend against their arguments. On the plausible assumption that many people find the intuition of responsibility misplaced after they have properly focused on the unavoidability point and the W-defence, it seems that the Frankfurt defender may have simply run up against the very intuition he was out to challenge in the first place, i.e. that alternative possibilities are necessary. Moreover, in terms of the overall dialectic here and the burdens of explanation each side bears, this looks to be more problematic for the Frankfurt defender than for Widerker and Otsuka, given the widespread acceptance of something like 'ought implies can', as opposed to the strength of intuitions in response to increasingly arcane counterexamples. As I said above in relation to Widerker, Otsuka's position here is likewise in harmony with my diagnosis of the structure of the Frankfurt cases anyway.

In conclusion, neither the classic nor any of the modified Frankfurt style cases we have examined (which represent the range of cases offered

in the literature), show that agents can be morally responsible for decisions for which they had no robust alternative possibilities. In addition to a comprehensive survey of the classic cases, I have offered new arguments as to why the buffer cases, despite their sophistication likewise fail to undermine the principle of alternative possibilities. However, this general conclusion about the Frankfurt strategy, although it means we need a leeway theory of moral responsibility, is not by itself sufficient to establish that this leeway theory must be incompatibilist. That is because there are also compatibilist leeway theories of moral responsibility. I therefore now turn to the evaluation of compatibilist attempts to capture the alternative possibilities condition on moral responsibility.

Chapter 2

2. Compatibilist theories of alternate possibility

The conclusion of chapter one is that Frankfurt style cases, when properly understood, don't show that alternate possibilities aren't required for moral responsibility. Furthermore, the discussion of Widerker's W-defence and Otsuka's related argument give us good reason to believe alternate possibilities are necessary for responsibility. Under the supposition that alternatives are necessary, in this second chapter I'll evaluate what I take to be the most prominent compatibilist theories of alternate possibility. Firstly, the so called 'traditional conditional compatibilism' or 'classical compatibilism' of Hume, G. E. Moore, R. E. Hobart and A. J. Ayer. Secondly, one of the more recent positions in the free will literature: the 'new dispositionalist' compatibilism developed by Kadri Vihvelin and Michael Fara which can be seen as a sophisticated advance on the old conditional compatibilism. Thirdly, I shall examine the new 'agentive modality' analysis of compatibilist alternatives developed by Christian List.

I shall argue that all these various forms of leeway compatibilism run into significant problems. Many of these points are familiar. But it is in the discussion of Christian List's recent work that I advance two new arguments against this sort of compatibilist position. In particular, I focus on his account of alternative possibilities. In the first argument, I aim to

show that even if List defends the reality of free will as a power over alternatives in his 'higher order property - special sciences' sense, this sense isn't connected up with moral responsibility in the way required to speak to the traditional problem of free will and responsibility. Secondly, I develop a thought experiment based on agents having access to microphysical information about the location of the particles that constitute their bodies to demonstrate that, List's model notwithstanding, there are still powerful incompatibilist intuitions that are hard to explain away. My conclusion is that although List's arguments are original and compelling in many respects, they also fail to establish compatibilism as a credible position.

2.1 Traditional conditional compatibilism

Hume puts forward the first recognisable conditional compatibilism or 'conditional analysis' of freedom in the *Enquiry Concerning Human Understanding*:

> by liberty, then, we can only mean a power of acting or not acting according to the determinations of the will – that is, if we choose to remain at rest, we may; if we choose to also move, we also may.[62]

The view was taken up and popularised in the twentieth century, most notably by G. E. Moore and A. J. Ayer. The basic idea is that ability claims in natural language such as 'I could do X' or 'I could have done X' are equivalent to subjunctive conditionals such as 'I would do X if I chose

[62] Hume, David. 1777. *Enquiry Concerning Human Understanding*, s. 73

to do X' or ' I would have done X if I had so chosen (or tried)'. It is then pointed out that the truth of such conditionals is consistent with determinism. Even if it's fixed by the past state of the universe and the laws of nature that you were determined to do X at a certain point in time, the conditional may still be true that if you had chosen to do something other than X at that time you would have done. It might be true or false depending on what you would have done if you had decided differently, even if you were determined to decide as you actually did. I take it to be uncontroversial that the truth of such subjunctive conditionals would not be undermined by the truth of determinism. The criticisms of this position therefore turn on whether the semantics of everyday 'coulds' and 'cans' really are correctly analysed as such conditionals.

Two of the most influential criticisms of the old conditional analysis were developed in the 1960s by Roderick Chisholm and Keith Lehrer. Take the following case of a man called Brown who fails to jump into the sea at some point in time to save a drowning child. In this case we can stipulate that the man is so terrified it is psychologically impossible for him to jump into the sea as he is. The conditional analysis claims that the following two claims are equivalent:

A. Brown could have jumped into the sea at t.

B. If Brown had chosen to jump into the sea at t, he would have jumped into the sea at t.

There are various formulations of the relevant subjunctive conditional in the literature with 'willed' or 'tried' sometimes being used instead of 'chosen'. These variations won't make any difference with respect to whether or not the following criticisms are salient. The problem is as follows. Suppose that, for whatever reason, it was *psychologically* impossible for Brown to jump in. Maybe he has a phobia of water, or he's terrified of drowning, or something similar. In such a case it is still seems true to claim that if he had chosen to jump he would have jumped (hence B can be true), yet it also seems false to claim that Brown could have jumped in – surely he couldn't given his phobia etc. Hence A seems false. This kind of case appears to show that A and B don't have the same truth conditions, and that consequently, they can't be equivalent.

There are other arguments in the contemporary literature that tell against the conditional analysis. Christian List has recently defended an unqualified (i.e. non-conditional) sense of 'could' as part of his compatibilist position. He first criticises the conditional analysis in the following way (List credits Keith Lehrer with the original argument behind this):

> A useful test for whether an interpretation of something is adequate is whether that interpretation can be substituted for its target without changing the meaning too significantly. To see that a conditional or dispositional interpretation of the ability to do otherwise fails this substitution test, notice that the sentences
>
> (1) the agent does not try to do X,
>
> and
>
> (2) if the agent does not try to do X, he or she cannot do X,

entail the negation of

(3) the agent can do X,

whereas they do not entail the negation of

(3*) if the agent were to try to do X, he or she would succeed in doing X.[63]

List points out that (1) and (2) appear consistent with (3*) whereas they are clearly inconsistent with (3). The upshot is that (3) and (3*) cannot be equivalent. The actual truth values of the above sentences are irrelevant in establishing this point.

The arguments above from Chisholm and List strongly suggest that the everyday sentences we form with 'coulds' and 'cans' that are partly constitutive of the ability claims at issue in the free will debate do not have the same truth conditions as simple subjunctive conditionals.

There is also the related problem that, upon encountering the conditional analysis, such as 'Agent A would have done X had they so chosen', people often ask (of the antecedent in the subjunctive conditional – in this case the 'had they so chosen'): but could the agent have so chosen? The critical thought behind this question is supposed to be that this is the relevant question at issue if we're worried about free will, and not the question of whether the subjunctive conditionals in the conditional analysis (that tie hypothetical (different) decisions to (different) bodily actions) are true.

---

[63] List (2014: 159-160)

Now, strictly speaking, this is a dialectically unsatisfactory response because the conditional compatibilist has presumably already argued for the conditional analysis. In this context (where arguments for the analysis are already on the table) an internal critique is properly called for (as in the case of Brown and the drowning child example above) or in any other form that would challenge the original argument the conditional analyst gave in favour of the analysis or its internal coherence. If this isn't done and the question about the antecedent is asked in isolation, the conditional compatibilist is entitled to reply that the question asked of the antecedent is appropriately also equivalent to another subjunctive conditional, in this case 'Agent A would have so chosen, if they had so willed/tried/chosen so to choose'. Presumably the interlocutor here will stubbornly go on to ask exactly the same question about the antecedent of this conditional analysis of the original antecedent. Worries about the intelligibility of 'chosen so to choose' to one side, the question (posed on its own without further argument) therefore doesn't properly engage dialectically with the conditional compatibilist, it's a mere expression of the fact that the analysis 'seems wrongheaded' to them or whatever. However, that said, the widespread insistence of a chain of questions of this sort suggests to me (even if they are not strictly speaking dialectically good arguments for the claim) that the ordinary language semantics of 'coulds', 'cans' and ability claims of the sort at issue in the free will debate are clearly not *intuitively* cashed out in terms of subjunctive conditionals.

At the very least, the conditional compatibilist owes us an explanation of why these regressive questions about the antecedents are so often asked.

There is however a better way to press the worry in the paragraph above (which I will term 'asking the could question about the antecedent of the conditional') that is both dialectically acceptable and problematic for the proponent of the conditional analysis. In the above paragraph I said that it might not make any sense to say 'would have chosen had they chosen so to chose'. This sentence is the one that is claimed to be equivalent to 'could have chosen', in the 'could question' about the antecedent which is pressed upon the conditional analysis. The worry now is that if it doesn't in fact make sense to say 'would have chosen had they chosen so to choose' then we will have a reductio ad absurdum argument against the conditional analysis based upon the claimed conditional equivalences that the proponent of the analysis presumably *must* give. I say presumably here because perhaps it is possible for the conditional analyst to claim that not every 'could' claim is equivalent to a conditional. However, if they do this we will be entitled to ask what the truth conditions of that 'could' are and whether they are consistent with the truth of determinism which was the whole motivation for the position in the first place. As the argument stands then, it looks like the conditional analyst will probably stick to their guns and offer the conditionalised version of the could question. What sense can we make of 'chosen so to choose'?

I think the interpretation of 'chosen so to choose' presents the conditional analyst with a dilemma depending on whether we take the sen-

tence literally or not. On the first horn, perhaps it is just a rhetorical way of saying nothing other and above that the person chose a certain way (perhaps with the rhetorical flavour here indicating careful or mindful choice etc). However, if that's all that's being (literally) said there will be a problem as the conditional will become trivial and uninformative, it will in fact read 'would have chosen X, if they had chosen X' which is a tautology. On the second horn, if 'chosen so to choose' is taken literally in a way that implies two choices have been made, it seems just plain wrongheaded to speak like that. You don't choose to make a certain choice – you just make a choice a certain way after deliberation. It is important to note that at this point in the argument, to say 'chosen so to choose' seems incoherent is not analogous to merely pressing the 'could question about the antecedent' as was described as dialectically unsatisfactory above. The claim now being made in the second horn is that the phrase seems incoherent, which is different to saying a position just seems wrong. The thought here is that it's difficult to make intelligible what (when one makes a choice) saying one chose to make that choice could mean over and above the fact that someone made a choice a certain way for certain reasons. There is another problem on the second horn as well. It looks like a regress threatens if every 'could question' is rendered as a conditional. If every time, the could question is asked about the antecedent of some conditional, it will be claimed equivalent to another conditional a regress will result. It's not clear to me that this is

necessarily a problem in itself, but it is the argumentative burden of the conditional analyst to explain why not.

Either way then, I claim that pressing the 'could question about the antecedent' enables a reductio argument to be made against the traditional conditional analysis.

2.2 The new dispositionalist analysis

Some criticisms of the traditional conditional analysis outlined above were on the whole considered decisive by compatibilists and incompatibilists alike during the rest of the twentieth century. However, the failure to secure compatibilism by this route had soon become less of an issue given Harry Frankfurt's development of the view that alternate possibilities are not required for moral responsibility. The main motivation to find a way of rendering the everyday modal claims about ability to do otherwise consistent with determinism was therefore (for many) deflated and it's fair to say that large numbers of compatibilists embraced Frankfurt's position and ceased to worry about the problems of the conditional analysis. Interestingly however, during the first decade of the twenty first century the core idea behind the conditional analysis has been revived and a number of prominent philosophers have defended a somewhat analogous position which has come to be known as the 'new dispositionalism'. In the rest of this chapter I will outline and evaluate whether the

new dispositionalist position does any better than its twentieth century relative.

Kadri Vihvelin (2004), Michael Fara (2008) and Michael Smith (2003) have developed the new dispositionalist position. Vihvelin argues that the old conditional compatibilism was on the right track with respect to the fact that it offered us an analysis of ability in terms of dispositions but that it was mistaken to then analyse the relevant dispositions as simple subjunctive conditionals. We have already discussed the problem Chisholm raises whereby the relevant subjunctive conditional can be true for some situation whereas the agent in question intuitively does not have the relevant ability that the conditional is meant to be equivalent to. This was illustrated by the man who couldn't jump in the sea to save the child, or in other examples, the girl that can't pick up the spider out of paralysing fear though *she would if she tried* to etc. Vihvelin claims that "persons have abilities by having intrinsic properties that are the causal basis of the ability."[64] She gives the following analysis of ability, which is very similar to David Lewis' revised analysis of dispositions themselves:

> **(RCAA)** *S* has the ability at time *t* to do *X* iff, for some intrinsic property or set of properties *B* that *S* has at *t*, for some time *t'* after *t*, if *S* chose (decided, intended, or tried) at *t* to do *X*, and *S* were to retain *B* until *t'*, *S*'s choosing (deciding, intending, or trying) to do *X* and *S*'s having of *B* would jointly be an *S*-complete cause of *S*'s doing *X*.[65]

[64] Vihvelin (2004: 438)

[65] Vihvelin (2004: 438)

RCAA stands for 'revised conditional analysis of ability', following David Lewis' 'revised account of dispositions' which is as follows:

> **(RCAD)** Something x is disposed at time t to give response r to stimulus s iff, for some intrinsic property B that x has at t, for some time t' after t, if x were to undergo stimulus s at time t and retain property B until t', s and x's having of B would jointly be an x-complete cause of x's giving response r.[66]

Lewis defines 'x-complete cause' in his RCAD as "a cause complete in so far as havings of properties intrinsic to x are concerned, though perhaps omitting some events extrinsic to x."[67] By analogy, I will understand Vihvelin's S-complete cause in her RCAA in just the same way: 'a cause complete in so far as havings of properties intrinsic to S are concerned, though perhaps omitting some events extrinsic to S'. This is what Clarke refers to as an 'agent-complete cause' in his criticisms of Vihvelin discussed below.

The above account is primarily intended as an analysis of basic abilities, abilities that are dispositions. More complex abilities, including the ability to make choices based on reasons, are not dispositions per say but bundles of dispositions.[68] RCAA is supposed to circumvent some of the problems of the old conditional analysis. But does it? The objection discussed above showed that the truth of the conditional analysis *analysans* (would have chosen if had tried to etc) was not sufficient for the posses-

---

[66] Lewis (1997: 157)

[67] Lewis (1997: 156)

[68] See Vihvelin (2004: 439)

sion of the ability itself. Randolph Clarke notes that this problem appears to apply to RCAA just as well:

> It might be true, as Vihvelin (2004, p. 443) says, that an agent (an animal, or a young child) can have an ability to perform an action of *A*-ing without having any ability to choose to *A*, but the further possibility that some agent might lack that ability to choose and, *for just that reason*, lack the ability to *A* undermines RCAA as an analysis of ability to act. A phobic agent might, on some occasion, be unable to choose to *A* and unable to *A* without so choosing, while retaining all that she would need to implement such a choice, should she make it. *Despite lacking the ability to choose to A, the agent might have some set of intrinsic properties B such that, if she chose to A and retained B, then her choosing to A and her having B would jointly be an agent-complete cause of her A-ing.*[69]

So just as before, it looks like the *analysans* is not sufficient for the truth of the unqualified ability to act. There was another criticism of the traditional conditional analysis which I didn't mention above that aims to show that the conditional *analysans* is in addition not necessary for the possession of the ability in question. J. L. Austin's example of the skilled golfer who just misses a short putt appears to be a case where the golfer *could* have holed it in spite of the fact that *he didn't hole it though he tried*.[70] Does an analogue of this apply to RCAA as well? Clarke says yes because the presence or absence of the 'causal basis of the ability' makes no difference here. Vihvelin's analysis gives the same result on the golf case as the original conditional analysis did.

---

[69] Clarke (2009: 329) My italics.

[70] Austin (1961)

I don't think the golf putt objection is a criticism of the relevant kind here though. That's because it seems that the 'could' in the sentence 'the golfer could have made that putt' is plausibly the 'could' of general ability and not the 'all in' narrow ability that we're after for free will. The sense of ability that's at issue in the free will debate is not general ability and so the golf example doesn't on the face of it give us an argument to say that Vihvelin's analysis is wrong here. The fact that (as the golf putt example shows) we can retain general abilities during periods and episodes where we lose the relevant 'all in' or narrow ability is widely accepted. The man who freezes at the moment when he should play his solo despite being a world class musician is a classic example of this kind of case. What we would need here is a golf case where the 'all in' or narrow ability to sink a putt was clearly the sense at issue in the phrase 'but he could have putted it'. I think it's problematic to imagine such a case because the environmental factors that can get in the way between the decision to strike the ball a certain way and the final position of the ball would seem to make that an inappropriate thing to say in the first place. This point can be appreciated by the comparison with the sense of 'could' in the other criticism that aimed to show the *analysans* wasn't sufficient. The 'could' in the sentence 'the man could have jumped in the sea' is much more plausibly the narrow 'all in' 'could' at issue here. We (arguably) mean something like - right there and then, holding everything else fixed, he could have jumped rather than not jumped - whereas this seems at best unlikely for the 'could' in the golf putt case. J. L. Austin

notoriously argued that we can, in certain contexts, mean the 'all in' or what I've also called the 'narrow' sense of 'could' in cases such as the golf putt but I will set this aside as I think there are more powerful objections to press here.

Fara's treatment of these issues is slightly different and is not immediately subject to analogous criticisms. He offers the following dispositional analysis of abilities to act:

> **(DAA)** An agent has the ability to *A* in circumstances *C* if and only if she has the disposition to *A* when, in circumstances *C*, she tries to *A*.[71]

He goes on to claim that if dispositions were given a simple conditional analysis, then his DAA would run into the same problems as the original conditional analysis. Consequently, Fara rejects that analysis and instead leaves dispositions unanalysed in his account of abilities. This might appear surprising but it isn't necessarily a problem if uncontroversial claims about dispositions are used to deliver and explain our intuitive expectations of what's required by unqualified abilities to act.

Fara thinks he can get around the problem of Austin's missed golf putt (the problem of the *analysans* not being necessary for the ability in question). According to Fara, DAA allows that the golfer *can* make the putt despite the fact he tried and failed because the golfer's disposition to make the putt when he tries can be subject to masking and consequently, so can his ability to make the putt. I won't evaluate this move in response to the golf putt because as I have said above, I don't think this

[71] Fara (2008: 848)

is the appropriate type of criticism for the incompatibilist to make here. I think there is a much more clear cut problem with Fara's proposal given the related problem of the analysans not being *sufficient* for the ability. Returning to the man who couldn't jump in to the sea and the child so scared of spiders that she couldn't even try to lift one, does Fara do any better then Vihvelin here? To reiterate, the problem was that it seems right to say that the child *can't* lift a spider, despite the fact that it might be true to say that they *would* lift one if they tried (on the old conditional analysis). The problem was that they couldn't try given how frightened they were. Analogously for the man who couldn't jump into the sea. Fara's response here is interesting as he says his DAA analysis would concur. He concedes that the girl can't lift the spider and that the man can't jump but this isn't a problem because the child *doesn't* have a disposition to lift the spider here anyway because her trying to lift a spider is impossible. Fara argues that objects don't have dispositions to do things in conditions that can't be realised. Clarke points out that Fara offers no analysis of the type of impossibility that is at issue here but we are given examples:

> if a rubber ball is nailed to the wall, and so cannot (in the relevant sense of "cannot") be dropped onto the floor, it is no more disposed to bounce when it is dropped than it is disposed to melt when it is dropped; it simply lacks any dispositions to behave one way or the other when it is placed in conditions that it cannot be placed in.[72]

[72] Fara (2008: 852)

I find this position very counterintuitive and agree with Clarke that that ball nailed to the wall *does* have a disposition to bounce if dropped here. The fact it can't be dropped (even ever) does not seem relevant to that claim. Clarke goes on to say:

> Likewise, it seems, salt that is permanently sealed in an unbreakable container remains soluble, disposed to dissolve if mixed with water, even though it cannot be so mixed. (If being so enclosed deprived it of solubility, that disposition would not be an intrinsic property.) Similarly, two particles might be disposed to interact with each other in a certain way despite the fact that they are so short-lived and so distant that they cannot be brought into interaction. (Our evidence that they have such a disposition might be the actual interactions of pairs of intrinsic duplicates of them.) If there is some kind of impossibility of manifestation conditions (short of metaphysical or nomological impossibility) that precludes possession of a given disposition, Fara has not made clear what it is.[73]

I think these comments are decisive but charitably, Clarke finds another way to press the sufficiency of the *analysans* worry, independent of the examples of not jumping into the sea and picking up spiders. He asks us to imagine a case where a man forms an intention to wave but then when the time comes, due to a temporary neural glitch, he can't even try to wave, though he would have waved had he managed to try. Does this man have a disposition to wave when he tries to wave asks Clarke? Clarke describes the following case to help sharpen the intuition that the answer to that question is yes. He describes a coffee dispensing machine which drops sugar into the coffee when the sugar button is pressed. Due to a temporary malfunction with the internal mechanism, the sugar but-

[73] Clarke (2009: 335)

ton is pressed and the sugar is not released. It still seems intuitively correct to describe the machine as retaining its disposition to drop sugar into coffee. Likewise with the man who couldn't wave due to his neural malfunction, he seems to retain the general disposition here. In conclusion, I take these examples from Clarke to show that there is a sense of ability, plausibly the one at issue in the free will debate, which is not captured by the new dispositionalist analyses I have discussed.

In conclusion then, it seems that criticisms analogous to those that apply to classic conditional compatibilism do after all apply to the new dispositionalist analyses as well.

2.3 A compatibilist 'could' that's neither conditional or dispositional: Christian List and 'agentive modality'

The attempts to give conditional and dispositional analyses of 'could have done otherwise', surveyed in the last two sections, are problematic and appear at odds with natural language semantics. However, there are still other options for compatibilists. An understandable move at this point might be to give up on alternatives altogether and embrace the Frankfurt style strategy. If it was a choice between abandoning compatibilism for free will and responsibility scepticism or dropping the alternative possibilities requirement (assuming that strategy is well motivated), many people would and have gone down the latter route. In the second half of the twentieth century, many compatibilists adopted the Frankfurt strategy given the problems that beset classical conditional compatibil-

ism. Assuming the new dispositionalist analyses don't provide us with a satisfactory treatment of alternative possibilities either must we turn away from the requirement? Christian List has recently argued that we can and should embrace a compatibilism where alternative possibilities are central to free agency. Moreover, he adopts an interpretation of 'could have done otherwise' where the alternate possibilities are unqualified modalities, as List also rejects classical and new dispositional compatibilism. List concedes that embracing a modal interpretation of alternate possibility in this unqualified way might look like a non starter for the compatibilist project, as it was precisely the tension between the fact that determinism entails there is only one physically possible future and the assumption of alternative possible futures that gives rise to the free will problem in the fist place. However, the unqualified modal interpretation of 'could' is the one List develops and defends. I will outline and evaluate List's argument below.

List ingeniously combines a number of claims which together enable him to argue for the position that physical determinism wouldn't deprive us of the variety of alternate possibility we need for free will. Before examining those, it will be useful to revisit the traditional argument for incompatibilism in order to set the context for List's point of departure. List gives the basic form of argument for leeway incompatibilism as follows:

> Premise 1: A necessary condition for someone's action to count as free is that the agent can do otherwise.

> Premise 2: Determinism implies that the agent cannot do otherwise.
>
> Conclusion: Either there are no free actions, or determinism is false (or both).[74]

The same basic syllogism underpins van Inwagen's consequence argument. List doesn't challenge the first premise, that has already been done by the Frankfurt example strategy and evaluated in chapter one. List focusses instead on the second premise. He argues that premise 2 is false even when we interpret the ability to do otherwise in an unqualified modal way and not conditionally or dispositionally as I said above. With this in mind List specifies the basic argument more precisely:

> Premise 1: Free will requires that (at the time of interest) more than one alternative course of action is possible for the agent.
>
> Premise 2: Determinism implies that (at the time of interest) only one alternative course of action is possible *for the agent*.
>
> Conclusion: Free will and determinism are incompatible.[75]

List says that the thesis of determinism is a claim about *physical* possibility. The truth of physical determinism doesn't entail the truth of the second premise here as that premise is about what's possible for an *agent*. Premise 2 above would be true if a certain linking assumption was true connecting physical and agential possibility. It would have to be the case that if at a given time only one sequence of events was physically possible then, as a consequence of that fact, only one course of events would be possible for the agent. List says that this is not generally true. This

---

[74] List (2014: 156)

[75] List (2014: 160) My italics.

claim, that the linking assumption is false is not question begging against the definition of determinism precisely because there is a difference between these two types of possibility, agential and physical, and therefore they can't be just assumed to run together.

What is List's argument for the distinct type of agentive modality being the salient one here then? His case is built on some basic methodological and theoretical assumptions from the philosophy of science and the special sciences. Essentially, the claim is that physical possibility is an inappropriate level of description (or frame of reference) when we are talking about what's possible for an agent;

> When we are interested in whether a particular action is possible for an agent, by contrast, the appropriate frame of reference is not the one given by fundamental physics, but rather the one given by our best theory of human agency. Thus the description of the world that matters here is not a (microscopic) physical one, but a (macroscopic) psychological one. Candidate theories that provide the right level of description include some advanced versions of psychological decision theory, such as those we find in economic psychology or cognitive science, which are currently our best attempts to make scientific sense of intentional agency. In fact, even folk psychology outperforms physics or neuroscience when it comes to understanding and explaining human behaviour across different domains and outside isolated laboratory conditions.[76]

What exactly is it though that makes the course grained 'macroscopic' reference frame the one we *should* be working in when we evaluate what's possible for an agent here? List appeals to the natural ontological attitude in addition to some claims about supervenience and multiple

[76] List (2014: 161-2)

realisability of mental states.[77] Together these claims support the distinction and the point that the agentive and not the physical modality is the relevant type of possibility when thinking about an agent's options. The natural ontological attitude is the idea that we should posit the theoretical commitments of our most successful theories as real, as existing in the world. For example, in fundamental physics, if electrons and quarks are part of the scientific model that allows us to most accurately explain what's going on in the physical world then electrons and quarks really exist. In a similar vein, when we are modelling and analysing human behaviour in the social sciences, the most successful theories are decision and social choice theory. Crucially, those theories posit the ability to select options from sets of possibilities on the basis of rational deliberation. That feature is part of the theoretical apparatus that allows us to best model and make sense of human behaviour. Analogously then, the reality of agents choosing options from sets of alternate possibilities is consequently brought about. This is what allows us to say free will exists just like electrons exist! In this instance, it's a higher level (special sciences) phenomenon - an ontology akin to desire and belief. List says that, "… Free will, in the technical sense of an agent's having a choice between more than one course of action in many situations, is a key presupposition of our best scientific theories of agency, at least when these theories

[77] On the natural ontological attitude see Quine, W. V. O. (1977) Ontological Relativity and Other Essays, New York: Columbia University Press and Fine, A. (1984) "The Natural Ontological Attitude", in J. Leplin (ed.), Scientific Realism, Berkeley: University of California Press, pp. 83-107.

are understood literally."[78] So it's because the agentive modalities are part of our most successful theories about human behaviour that we should focus on those modalities, rather than on (microscopic) physical possibility. The microphysical scientific reference frame is not one where we have anything like success in modelling human behaviour. It is virtually impossible to explain even basic human behaviour in microphysical terms. You need a supercomputer, Laplace's demon or God and even then the content of those physical statements wouldn't obviously translate into understandable propositions in terms of agency as there are additional worries about whether one can map and translate higher-level descriptions from microphysical ones. This is what underwrites List's claim that we should work with agentive, not microphysical modality. I will not here criticise anything List says about multiple realisability and the supervenience of agential states on the microphysical states of the brain. I wish to grant all the latter for the purposes of argument. Likewise with the toy model List develops to demonstrate how the agentive modalities might supervene on the physical modalities. I agree List has successfully shown how *in principle* higher level indeterminism can supervene on a system which is deterministic at a lower level of description or frame of reference.[79] The key question however, is whether the semantics of that higher level agential indeterminism List argues for are what we need to speak to the traditional free problem. I shall focus on two is-

[78] List (2014: 168)

[79] See List (2014: 162) for the toy supervenience model

sues here. Firstly, I will examine the status of the free will List claims exists given our best theories of human behaviour and folk psychology and develop a new line of argument that shows it doesn't follow that *this* kind of free will is appropriately connected up with the conditions on moral responsibility, as any concept of free will must be if it is to be relevant to the core issue at the heart of the traditional problem of free will and moral responsibility. That is to say there is a gap in the argument here that must be plugged. Secondly, I will develop a new argument that directly challenges the claim that the agentive modalities as List explicates and defends them are coextensive with our ordinary language concepts of free will (and consequently moral responsibility). Specifically, I will develop and outline a case where, in spite of what List claims, facts about the microphysical level of description *are* intuitively pertinent and constraining on what can be said to be possible for agents at the psychological course grained level of description. If my second argument is sound then the incompatibilist will be in a position to not only claim that List's 'agentive possibility' is not necessarily our ordinary language concept at issue in the free will debate, but in addition, that our everyday concept of free will is still incompatible with microphysical determinism. I now turn to these two arguments.

2.4 First response to List: The ontology of free will and the link with moral responsibility

I now wish to examine in more detail what the metaphysical status of free will arrived at in the way List sets out amounts to. On this account, we can't view free will merely instrumentally – as a useful tool that doesn't have any ontological significance (an analogue of the instrumental view in the philosophy of science). List specifically rules this out because he thinks the natural ontological attitude forces us to take our commitment to agentive modalities at 'face value'. Instrumentalism of this kind would be another 'ontological option' here if we're considering the analogous possibilities from the philosophy of the physical sciences. However, once again, I won't challenge this point for the purposes of argument as what I want to say still applies when we do embrace the special science ontology of freedom in the way List wants us to.

The first thing to note is that the ontology of free will generated by the application of the naturalistic ontological attitude to the special sciences and social choice theory need be no thicker than the theories making use of the term *require in order that they constitute successful theories.* Given that rational choice theory is trying to model and explain why we act as we do and how we should act given our desires and goals etc, it seems uncontentious that the modal notions it makes use of need not be those of (or make reference to) microphysical possibility, instead course grained phycological descriptions are sufficient here. Microphysical descriptions are in fact inappropriate, not merely unnecessary. To say this is not to

make the instrumental move either. The point is rather that the relevant entities such as 'choice over option sets' can enable the theories in question to work to their full potential without needing to be metaphysically anything over and above an agential-psychological phenomenon. Specifically, we don't need the notion of 'choice over alternatives' that's in play when we're working in social choice theory to require the truth of *physical* incompatibilism. If we're trying to model the *rationality* of behaviour, then we can do it successfully with a thinner agential (compatibilist) notion of choice over alternatives.

With all this in place I can proceed to give my arguments against List. These are motivated by questioning whether his sense of free will is: (i) the same as our ordinary language sense, (ii) what we want, or, (iii) the one we need? In short, is List's free will the free will that's traditionally *at issue* in the free will debate?

A worry — what are the ordinary language semantics of 'can' or 'could have done otherwise'?

I wish to press the point that in order for List's argument to constitute a successful defence of compatibilism about our concepts of free will and responsibility, he must go on to demonstrate that the truth conditions of the natural ontological attitude (NOA) generated sense of free will are coextensive with our ordinary language sense. The key point here is that it doesn't follow from the fact that we can derive an ontology of free will in the way that List describes that *that* sense of free will is the same as

our ordinary language notion of free will, i.e. the sense at issue in the free will debate. That is, it doesn't follow from his NOA argument that the everyday sense in which we think agents can do otherwise than they do is underwritten by the compatibilist modal semantics that List develops in his toy model. This is because List's central argument is not an analysis of the semantics of ordinary language terms but is instead built around the deployment of a theoretical (methodological) device, namely the NOA, which tells us what we *should* and shouldn't have in our ontology given our best theories of agency. I don't wish to take issue here with the use of the NOA, I'm instead claiming that there is a gap in the argument unless we are shown that List's free will is the same as the ordinary language concept.

That said, List does very briefly discuss the ordinary language issue in section six of the paper. He notes that the present analysis is broadly consistent with some accounts of the ordinary meaning of 'can.' He then goes through some examples from Angelika Kratzer's account from the 1970s, and mentions work by Ann Whittle and John Maier on the agentive modalities. In response, my point is simply that none of these arguments are uncontentious. A much more thorough analysis is required. Perhaps it might be said that because List has put a compatibilist modal semantics on the table, the ball is in the incompatibilist critic's court given the burden of proof in the dialectic. I concede that this would indeed be the case if List's modal semantics had been derived from ordinary language analysis but they're not. Furthermore, as I shall demonstrate in

the second argument below, we have good reason to think that the 'agentive' modal semantics List develops are not coextensive with our ordinary language usage.

At this point in the discussion it is worth remembering the distinction between the two types of compatibilism I outlined in the introduction: diagnostic and prescriptive. It seems that List isn't exactly clear which of these camps (or both) his project falls into. On the one hand the methodology of using the NOA might be seen as falling under the prescriptive heading. List's telling us we have *reason to adopt* a certain sense of free will independent of the conceptual analysis of ordinary language. On the other hand he does engage in diagnostic work here, precisely because he mentions the Kratzer semantics and is at pains to point out this is evidence that his analysis accords with ordinary usage. Towards the end of the paper and pertinent to this issue, List discusses what we should do if it was discovered that determinism were true and we woke up to the morning papers with headlines announcing the fact. In that circumstance, after considering options like ceasing with responsibility practices and the institution of punishment or in fact carrying on much as before, he says;

> Surely we would do the latter: giving up our conventional understanding of free will and revising the very fabric of how human society works would be an overreaction. The approach to free will offered in this paper shows why this is so. The mildest revision of our technical vocabulary—namely the shift from physical to agential possibility in the analysis of free will—is sufficient to rehabilitate practically everything we conventionally believe and say about free will, even

> against the background of determinism. For this reason, my proposal seems to have common sense on its side.
>
> In conclusion, I suggest that the best way to defend the compatibility of free will and determinism is to recognize that free will is not a physical phenomenon, but a higher-level phenomenon on a par with other familiar higher-level phenomena such as beliefs, desires, and intentions. If we are searching for free will at the level of fundamental physics, we are simply searching in the wrong place.[80]

The phrase "mildest revision to our technical vocabulary" is telling here. Revision would imply that List's concept of free will is not exactly the same as the one we have been using? If he is happy to accept that the concept he advocates isn't quite the same as we find in ordinary practice, why is it so important to discuss the Kratzer semantics? Still, in defence of the latter discussion, presumably the milder the revision the better here and so just because a project is revisionist that doesn't mean that concern with ordinary usage evaporates. However, even if such revision is mild, we need to know whether the revision licenses the continued deployment of the family of concepts and practices that free will is taken to be connected up with in our ordinary beliefs and practices here. In short, we must discuss the link with responsibility.

The central issue in the free will debate is whether in a deterministic universe we would have the type of control relevant to moral responsibility. My claim is that after the application of the natural ontological attitude to rational choice theory, List still needs an argument to establish

[80] List (2014: 174)

that his resultant ontology of free will is sufficient for moral responsibility.

Traditional compatibilist analyses of 'could have done otherwise' in ordinary language automatically resolve this worry about responsibility (if they are successful). This is because the ordinary language concept of moral responsibility is linked to our ability to do otherwise in such a fashion that if we can be shown to have the latter (in a deterministic world) then we may (under the right conditions) be justified in predicating the concept of responsibility of particular actions and people. In ordinary language analysis, if we have shown something about free will then we have also at the same time shown something about a necessary condition of moral responsibility. But, in close analogy with what I have said above, the free will List has argued for must still be shown to be sufficient for moral responsibility because it's not obvious that it's the same as that necessary condition of responsibility deployed in ordinary discourse.

Interestingly this worry remains even for those who are not particularly concerned about whether List's free will is coextensive with our ordinary language sense of free will. Why? This is because you still have to show that List's freedom is sufficient to hold someone responsible (even if you're a revisionist about the concept of responsibility as well!) Imagine for example, a person very impressed with the NOA argument, who concedes that List has shown us a robust sense in which we can be said to have free will. Such a person may (indeed must) still intelligently

question whether it makes sense to hold people morally responsible for their actions in a *physically* deterministic world.

The reason why there is room for me to make this criticism independently of the worry about ordinary language semantics is because the special science theories that posit free will in List's sense don't obviously try to model our practice of holding agents morally responsible. Hence more argument is needed to establish that List's ontology of freedom should stop people worrying about the threat determinism might pose to moral responsibility. What rational choice theory *does* is model the rationality of agents given certain goals and desires they have. It models how we should act given our desires, resources, abilities etc. But why should we expect that the theoretic entities that special science theory commits us to would also be related in the right sort of way to our notion of moral responsibility? The problem is that those models, i.e. social choice theory, don't seem to make any assumptions, one way or another, regarding whether the sense of choice at issue in them is also the kind of choice that would be sufficient for the control condition on moral responsibility. In summary then, it may well be true that decision theory uses a notion of choice that commits us to the reality of alternate possibilities. In that sense List may well be right that we can be said to be free, or have free will. All of this said, it does not follow that the notion of freedom or free will that is relevant to these successful social science theories that List bases his argument on is the notion of free will that is relevant to the free will debate. List needs to give us an answer to these questions.

Note that I'm not saying that ordinary language analysis is crucial for any successful argument in this domain. The same point as I develop above must also be answered by those who advocate a revised conception of moral responsibility. Another way of proceeding would be to establish that the free will generated by List is connected up in the right kind of way with our practice of moral responsibility as we *should* conceive it (even if that too is a revised concept). That is, if it could be established that we should embrace control requirements for moral responsibility and it was true that those requirements were instantiated by List's account of free will, that might be a way of vindicating List's position.

2.5 Second response to List: Microphysical modality is relevant to assessing claims about what agents can and can't do

So far I have tried to develop an immanent critique of List's position by showing that even if we grant him his set up and resultant ontology of free will, this might not be what we were after or in fact need given our responsibility practices. I now wish to develop a different line of criticism concerning the relations between the physical and agentive modalities. According to List, in a deterministic world where there is only one physically possible future, it can be nevertheless true to say that an *agent* can do otherwise. Leaving aside the arguments for this position and their shortcomings (that was the focus of the first argument above), I will consider some consequences that microphysical facts appear to have on

what we intuitively take ourselves (described at the agential frame of reference) to be capable of.

Suppose I'm at the counter in the canteen and I'm trying to decide between chocolate cake and fruit salad. It will be true, barring any funny business, to say that I could have taken the fruit salad when I in fact take the chocolate cake. The 'could' in the last sentence is the 'could' of agentive possibility. But we may wonder about the microphysical bases on which these different 'agential' choices are supposed to supervene. On List's model, What about the physical supervenience bases of these different choices? On List's model, the supervenience base for the counterfactual choice of taking the fruit salad was physically impossible, given that the chocolate cake supervenience base was determined to be the case instead. This is because when we are evaluating what's possible at the micro-physical level, it is physical possibility that is relevant, not agential. So, given the distinction between the different types of possibility here List is bound to say that you could have chosen the fruit salad despite it also being the case that the physical base that would have realised that alternative decision was (physically) impossible. Is this problematic? To insist it is (without argument) is to simply beg the question against List's position here. After all, List has argued for that distinction and given reasons why it is the agential perspective or frame of reference that is the relevant one here. He has also explained how the different types of possibility are related in his toy supervenience model and hence why they needn't run together here. Nevertheless it seems legitimate to say

that his position is very counterintuitive. It is surely natural to think that when I make claims about what I can do at the agential level, on the assumption that I'm a physicalist, I am also committed to the supervenience bases of these different choices being *physically possible*. List seems to deny this and it is hard to understand how that can be right.

It might well be objected here that all my complaint amounts to is that List's position is counterintuitive. And, perhaps, in pointing to this as a problem I am simply demonstrating that I am still in the grip of the very modal confusion List was addressing. On the other hand, it does seem like the most natural thing in the world to insist that the physical supervenience bases for my agentive possibilities must be physically possible whenever the agentive possibility is there. Again though, this might just be a statement of the confusion when we talk in the *same sentence across reference frames* here. If something like that is right, at the very least one would welcome some more explanation from List as to why this common modal assumption is made, why it seems so natural and what exactly is mistaken about it. Perhaps he might say that have a deep (and yet mistaken) implicit view that the microscopic physical frame of reference is somehow 'more real' than the agential, or something of this kind, and maybe that may well be right. However, List has not provided enough to dispel this very natural line of response here.

In the mean time, I think it is possible to strengthen the objection that List's position is counterintuitive and argue that there is a substantial problem here. On the other hand I will now go on to argue that there is a

real problem here. I present this below in a new line of argument which concerns the locations in space time of the microphysical particles that make up our bodies. Even then, this argument may further illustrate that i am still in the grip of the modal confusion List claims is at the heart of incompatibilist thinking. But the argument may also demonstrate that List's position is not convincing.

The 'physical information machine' argument

As outlined above, List claims that the correct frame of reference for assessing claims about agential possibilities is not the microphysical one and so the truth of determinism would not mean agents couldn't do other than they in fact do even if microphysical determinism is true. My argument is intended to put pressure on that claim. I have already said that it seems weird to be committed to saying that an agent can do otherwise (all described with reference to the agential frame of reference) and yet hold at the same time that the physical supervenience bases of those agential alternatives are *physically* impossible (given the current state of the world and the laws etc). I will now develop this point further by means of an example.

Imagine that you are trying to decide whether or not to go and visit your friend at the weekend in Manchester. You live in London and, at the agential frame of reference there is nothing that would prevent this. You have the money, you would be welcome, you're not so unwell that you couldn't travel to Manchester etc. You can also stay in London if you so

choose. At the course grained agential frame of reference there is nothing stoping you either way. It's Friday night and you are wondering what to do in the morning, specifically wondering whether you should go or not. You think (correctly, according to List's model) that you *could* visit Manchester or you could stay. In this case imagined the crucial modal claim you make on Friday night is that 'I could go to Manchester'. Now suppose that physical determinism is true at your world and you don't end up going, and instead stay in London to work. Now, you also just so happen to have access to a 'physical information machine'. This machine can't communicate in anything other than the microphysical reference frame but it can print out a perfectly accurate set of microphysical coordinates for where the microphysical particles that constitute your body will be in space time at any point in time. Imagine that on Friday, you request a print out telling you where your set of particles will be tomorrow. The machine prints out a form that tells you the particles in your body will be in London all weekend. (We can stipulate into the example that the machine has the means to represent the area of space time we call "London" and its space time coordinates in relation to that part of space time we label as "Manchester" etc.) According to List this microphysical fact (its truth arising from the way in which the single line of physical possibility is playing out at your determined world) does *not* falsify the agential claim you make on Friday night when you say "I could go to Manchester tomorrow". But that now seems manifestly false: I claim that it seems that it does exactly that - how could you go to Man-

chester if the atoms that make up your body were always going to be in London at the weekend? Physical information about where your body is going to be in space time (described physically) *is* relevant to agential claims. You can't (at least in this life) go somewhere without your body in tow. If this is correct then it is not the case that determinism at the microphysical frame of reference is irrelevant to modal claims at the agential frame.

The above thought experiment is generalisable to any decision an agent is going to make where they think they are choosing from a range of options. This thought experiment seems to constitute a strong case for saying that our ordinary language claims about what it's possible for us to do are threatened by the truth of microphysical determinism.

What could be said in response to this thought experiment here? It might be said that my machine is not an innocent stipulation, given the problems of formulating bridge laws between subvenient and supervenient reference frames (as Fodor discusses) and that consequently I can't assume that the information on the print out would be interpretable in such a way that you would be able to deduce 'my atoms will be in London tomorrow' etc.[81] In response to this worry we can imagine a possible world as follows. Each microphysical particle that exists (and that is subject to the strict deterministic microphysical laws) has been uniquely la-

[81] see in particular: Fodor. (1974), "Special sciences (or: The disunity of science as a working hypothesis)", *Synthese* 28 (2), pp. 97–115, and Putnam. (1975) "Philosophy and our mental life", in *Mind, Language and Reality* (Cambridge: C.U.P.

belled by God. God has made the world including the machine. The machine has a microphysical GPS that can tell where the machine is situated in space time. It also has an 'airport style' body scanner that can scan your body and detect the uniquely labelled atoms in your body. Given the machine is programmed to know the initial starting state of the world for each numbered particle and the future coordinates of each numbered particle given the laws, the machine can tell you that the atoms labelled in your body (the ones just scanned in the machine) will not be more than 10 miles from where the machine currently is for the next 24 hours. If the machine is in London then you know your body is not going to be in Manchester tomorrow. It has been determined from the dawn of time that the atoms that constitute your body will be no more than 10 miles from the machine (which is in London). However, List's account has it that consistent with all this, you can go to Manchester tomorrow. That seems wholly implausible.

Perhaps it might be said that I am still equivocating between the agential and physical frames here. It is one thing to say that the particles constitutive of my body described purely physically can't be in a certain part of space tomorrow but that doesn't mean that *I can't go to Manchester*. The latter sentence is explicitly not to be assessed in terms of microphysical possibility according to List. Is the very equivocation List was outlining still implicit in my argument here then? In response, it's not obvious *why* we should think that my given my example is explicitly making a distinction between the different frames and kind of information they can

provide. Given certain physical information, it seems I can make deductions about what is possible for me as an agent in the world described non-scientifically. List needs to explain away this powerful intuition. It is difficult to see how he can without simply arguing that we *should* be only using agential possibility here. But that is not the point at issue in the traditional *diagnostic* free will debate. Hard-Incompatibilists may well agree with him that we should understand agential possibility in the way List outlines but the thought experiment suggests that we do in fact think microphysical properties (in this case spatial location) are relevant to assessing claims about what agents can do.

There is another related argument here that is simpler and less prone to worries about how the machine could talk in terms that we could 'read off' properties like 'in London' or 'in Manchester' from. Imagine the same case where you decide to stay in London rather than go to Manchester at the weekend. On Monday you learn that microphysical determinism is true. Armed with this knowledge you now know that the atoms that make up your body couldn't have been anywhere other than they have been in space time in the past. Worries about spatio-temporal information bridging aside, you now know for sure (i) that your constitutive atoms haven't been to that part of space time we call "Manchester' at the weekend. But now we also know (due to determinism), (ii) that those atoms couldn't have been there and so we again arrive at the claim, (iii) If I can't go to places that my atoms can't go to then I couldn't have gone to Manchester.

In conclusion, I have argued for two claims in this section. Firstly I have shown that even granting that List can demonstrate we do posses free will in the sense of being able to chose from among agential possibilities, it is not clear, and it does not follow from List's arguments that this sense is *the sense* of free will at issue in the traditional debate, i.e. the sense that we standardly conceive of as the control condition relevant to moral responsibility. If so, then in spite of his original and ingenious special science methodological argument and his formal supervenience model, List's 'free will' doesn't (without further argument) appropriately speak to the free will problem. Secondly, I argued that our intuitive understanding of the relationship between ourselves as agents and ourselves as physical entities composed of the particles that are the subject matter of microscopic physics puts further pressure on List's claims and is evidence that we cannot combine microphysical determinism with agential indeterminism in the way necessary to constitute a successful defence of compatibilism.

Chapter 3

3. The threat to compatibilism independent of the issue of alternative possibilities

3.1 Can agents be the 'source' of their actions in the sense required for moral responsibility if determinism is true?

In the first two chapters of this thesis I examined issues concerning alternate possibilities and moral responsibility. The first chapter was devoted to arguments that tried to establish alternatives weren't required for moral responsibility and the second chapter examined arguments to the effect that the kind of alternative possibilities required for responsibility are in fact consistent with the truth of determinism. I argued that both these different projects fail. However, even if either of these arguments had been successful (i.e. managed to put to rest the leeway worry), any resulting compatibilist theory would still have to face up to a different challenge from the incompatibilist: the *source* worry. The relationship between determinism and the source and leeway worries is not as it might at first appear. For instance, it might be thought that if alternative possibilities were not required then the purported threat of determinism to freedom and responsibility is at once removed. After all, if we don't need any 'wiggle room' in the actual sequence why would we require indeterminism? However, it doesn't follow that determinism is no threat even

if we don't require alternatives. The source worry is about whether we can have the kind of control relevant to moral responsibility if the causes of our actions ultimately trace back to factors and forces beyond our control. This last sentence is of course vague and contains words subject to both compatibilist and incompatibilist readings (for example; 'control' and 'ultimate' here are crucially at issue in the debate). Those are the issues at the heart of the source worry. Nevertheless, the point now is that these worries can still be pertinent for someone who has rejected the alternative possibilities requirement. On the contrary, a similar burden lies with someone adopting a compatibilist reading of alternative possibility. Armed with a sense of 'could' consistent with determinism and a good argument for it still isn't obviously to have the means to immediately reject the source worry. In both cases, it still makes sense to ask whether determinism would rob you of the control salient to responsibility if all your actions were traceable to events in the distant past coupled with the laws of nature. This worry can be motivated regardless of your position on alternative possibilities. Pereboom is a good example of someone who has rejected the need for alternatives as a result of the Frankfurt strategy (his buffer cases) but who maintains incompatibilism as a result of the source worry. Although it is fair to say that the most immediate and widely discussed issue linked with the threat of determinism is the problem of alternative possibilities, the free will literature also contains a broad family of related arguments that make determinism a threat without mentioning alternatives at all. The so called *direct* arguments, *ultima-*

cy arguments and *manipulation* arguments all proceed in this way. Direct arguments run roughly as follows: the laws of nature and the past states of the world jointly determine the current state of the world including your actions. You are not responsible for the past states nor the laws, therefore you are not responsible for the current state including your actions. See Peter Van Inwagen (1983) for the classic formalised version of the direct argument. Ultimacy arguments make use of the same core worry about source-hood. See Robert Kane (1996), Martha Klein (1990) and Galen Strawson (1986, 2010) for examples of ultimacy arguments. Klein captures the main idea when she says, "… Agents should be ultimately responsible for their morally relevant decisions or choices —"ultimately" in the sense that nothing for which they are not also responsible should be the source [or cause] of their decisions or choices." (1990, p. 51). Galen Strawson develops arguments based on a very demanding idea of being ultimately responsible using a notion of U - freedom (see in particular his 2010, Appendix G, p 291). Manipulation arguments on the other hand proceed by arguing that there is no relevant difference between a case of a manipulated agent (where we have a clear intuition of non responsibility due to the agent not being in control) and everyday normal deterministic agents (hence we should conclude people are not responsible if determinism is true). Pereboom (2001, 2014) had developed a detailed multi stage manipulation argument to this effect. In summary then, it should be clear that embracing the Frankfurt strategy, or on the other hand adopting a form of leeway compatibilism does not necessari-

ly give you the means to reject a direct, ultimacy or manipulation argument. All these arguments must be addressed on their own terms which don't mention alternative possibilities at all.

I will not discuss direct or ultimacy arguments in this thesis and instead focus on manipulation arguments designed to show that if determinism is true we can't be the source of our actions in the way required for moral responsibility. The most sophisticated manipulation argument in the literature is Derk Pereboom's four case argument. I will outline and evaluate the four case argument in the rest of this chapter and discuss how compatibilists might respond to it. The four case argument is difficult to evaluate properly in so far as assessing it requires carefully situating it against the background dialectical burdens between incompatibilist and compatibilist. Once these background dialectical issues have been made explicit it should be easier to assess the status of the intuitive reactions to the different cases Pereboom gives us but without such careful background stage setting the four case argument is apt to result in a quick stalemate with both sides in the debate running the cases in the direction that suits their position. This diagnosis will hopefully be borne out in the following discussion of the argument.

The key to any manipulation case is to coherently describe a situation where all the compatibilist conditions on moral responsibility are satisfied and yet intuitively the agent is not morally responsible. It's therefore important that the part of the story that elicits the intuition of non responsibility does not at the same time undercut basic conditions of

agency in such a way that the agent in question is immediately disqualified from being a candidate for performing intentional voluntary action. If that were the case then rather than being a specific challenge to a compatibilist theory we would simply have a description of non agency and the predictable intuition of non responsibility to go along with it. However, if a manipulation case is made coherent then the compatibilist must either respond by modifying their account and pointing out that in fact there is another condition on responsibility which is violated in the example or, alternatively, by biting the bullet and arguing that the agent is in fact responsible, contrary to intuition. Such a case as this, if made coherent, would constitute a simple challenge to the sufficiency of a set of compatibilist conditions on moral responsibility but it would not on its own be an argument to the effect that determinism rules out the relevant type of control necessary for moral responsibility. Consequently, Pereboom's four case argument starts with a manipulation case as described above (case 1) and then proceeds to describe other cases, which by degree become the ordinary deterministic world (case 4) in which agents go about their business in the normal way without any manipulation or other interference. Pereboom then argues that there is no relevant difference between the early cases involving manipulation and the later ordinary deterministic cases cases that would allow us to conclude that while the agents in the early cases were not responsible the agents in a normal deterministic situation are. Specifically, he argues that the best explanation of the non-responsibility intuition in the early cases (1 and 2) was the

fact that the agent's actions were causally determined by factors beyond their control and that we should therefore apply this same reasoning to the normal deterministic cases (3 and 4) and conclude that determinism rules out moral responsibility. The four case argument is outlined below.

3.2 Pereboom's four-case manipulation argument

All of the four cases involve Professor Plum deliberating and then deciding to kill White. The first case is supposed to be a scenario where it is very intuitive (to both incompatibilists and compatibilists alike) that Plum is not responsible:

> ***Case 1***: A team of neuroscientists has the ability to manipulate Plum's neural states at any time by radio-like technology. In this particular case, they do so by pressing a button just before he begins to reason about his situation, which they know will produce in him a neural state that realizes a strongly egoistic reasoning process, which the neuroscientists know will deterministically result in his decision to kill White. Plum would not have killed White had the neuroscientists not intervened, since his reasoning would then not have been sufficiently egoistic to produce this decision. But at the same time, Plum's effective first-order desire to kill White conforms to his second-order desires. In addition, his process of deliberation from which the decision results is reasons-responsive; in particular, this type of process would have resulted in Plum's refraining from deciding to kill White in certain situations in which his reasons were different. His reasoning is consistent with his character because it is frequently egoistic and sometimes strongly so. Still, it is not in general exclusively egoistic, because he sometimes successfully regulates his behavior by moral reasons, especially when the egoistic reasons are relatively weak. Plum is also not constrained to act as he does, for he does not act because of an irresistible desire – the neuroscientists do not induce a desire of this sort.[82]

[82] Pereboom (2014: 76-7)

Pereboom is quick to claim about case 1 that Plum meets all the standard conditions familiar from the history of compatibilism. For example, given Hume's requirements, the action is not out of character for Plum and the motivating desire is not irresistible. Similarly, with respect to Frankfurt's own hierarchical view, the effective desire to murder White conforms to the higher order desires Plum has in the right kind of way. Plum also satisfies the conditions for reasons-responsiveness in Fischer and Ravizza's theory as the relevant desires can still be be conditioned by the rational consideration of reasons for action. To satisfy R. Jay Wallace, Plum can regulate his behaviour by consideration of moral reasons. Furthermore, Plum has the ability to reflectively revise and develop his moral character over time, given the commitments of Mele and Haji. Plum allegedly meets these conditions in *all four* of the different cases in Pereboom's argument.

I find it intuitive that Plum is not morally responsible for killing White in case 1. Pereboom points out that one possible explanation for this is that Plum's decision is causally determined by the scientists and hence beyond his control. What about case 2?

> ***Case 2***: Plum is just like an ordinary human being, except that a team of neuroscientists programmed him at the beginning of his life so that his reasoning is often but not always egoistic (as in Case 1), and at times strongly so, with the intended consequence that in his current circumstances he is causally determined to engage in the egoistic reasons-responsive process of deliberation and to have the set of first and second-order desires that result in his decision to kill White. Plum has the general ability to regulate his actions by moral reasons, but in his circumstances, due to the strongly egoistic nature of his deliberative reasoning, he is causally determined to make his decision to

> kill. Yet he does not decide as he does because of an irresistible desire. The neural realization of his reasoning process and of his decision is exactly the same as it is in Case 1 (although their causal histories are different).[83]

I think it's also intuitive that Plum is not morally responsible in this second case as well. In case 2, the manipulation is in the pre-programming so that the relevant piece of egoistic reasoning occurs at the right time. No intervention in Plum's neural processes occurs during the deliberation as in case 1. But it's surely natural to think that the time difference doesn't matter, how could such a time lag matter with respect to the question of whether or not the agent had the kind of control required? Apart from the time lag the actual sequence plays out in exactly the same way in both cases 1 and 2. Again, one obvious candidate explanation for why Plum in case 2 is not responsible is that his action was causally determined by factors beyond his control. Both in cases 1 and 2, the instances of manipulation - local and remote - causally determine that Plum decides to kill White.

Cases 3 and 4 move away from manipulation altogether and describe more familiar scenarios. The role of these cases is not to illicit the intuition of non responsibility (that is precisely is what is contested in these cases between compatibilist and incompatibilist) but instead to make it clear that if a compatibilist wishes to hold Plum responsible in cases 3 and 4 they better be able to draw a principled distinction between these

[83] Pereboom (2014: 77)

scenarios and cases 1 and 2. This is exactly what Pereboom claims cannot be done.

> ***Case 3***: Plum is an ordinary human being, except that the training practices of his community causally determined the nature of his deliberative reasoning processes so that they are frequently but not exclusively rationally egoistic (the resulting nature of his deliberative reasoning processes are exactly as they are in Cases 1 and 2). This training was completed before he developed the ability to prevent or alter these practices. Due to the aspect of his character produced by this training, in his present circumstances he is causally determined to engage in the strongly egoistic reasons-responsive process of deliberation and to have the first and second-order desires that issue in his decision to kill White. While Plum does have the general ability to regulate his behavior by moral reasons, in virtue of this aspect of his character and his circumstances he is causally determined to make his immoral decision, although he does not decide as he does due to an irresistible desire. The neural realization of his deliberative reasoning process and of the decision is just as it is in Cases 1 and 2.[84]

In order for the compatibilist to claim Plum is responsible in case 3, they must explain what necessary condition on responsibility is not met in cases 1 and 2 but is met in case 3. Pereboom claims it is difficult to see what this might be. Finally we have a 'normal deterministic world' scenario with no manipulation, funny business or problematic upbringing for Plum:

> ***Case 4***: Everything that happens in our universe is causally determined by virtue of its past states together with the laws of nature. Plum is an ordinary human being, raised in normal circumstances, and again his reasoning processes are frequently but not exclusively egoistic, and sometimes strongly so (as in Cases 1–3). His decision to kill White issues from his strongly egoistic but reasons-responsive process of deliberation, and he has the specified first and second-order desires. The neural realization of Plum's reasoning process and

[84] Pereboom (2014: 78)

> decision is exactly as it is in Cases 1–3; he has the general ability to grasp, apply, and regulate his actions by moral reasons, and it is not because of an irresistible desire that he decides to kill.[85]

There seems even less scope for making a distinction between cases 3 and 4 on the basis of the violation of some necessary condition on responsibility. Hence we appear compelled to treat cases alike and conclude that Plum is not morally responsible in case 4. This completes Pereboom's argument as case 4 is just the standard deterministic case. In summary, it's intuitive that Plum is not responsible in case 1 and there are no salient differences between cases 1 and 2, 2 and 3, and finally 3 and 4 that allow us to explain why Plum wouldn't be responsible in the first case in one of these pairings but not the second. Moving through the cases from 1 to 4, we should therefore conclude that Plum is not responsible in case 4. That is the four case manipulation argument for source incompatibilism.

3.3 Criticism of the four-case argument

I will now evaluate the four case argument paying careful attention to recent criticisms of it made by Alfred Mele, Michael McKenna and Kadri Vihvelin. As a first pass at how the compatibilist might try to drive a wedge in between the cases in a principled way, William Lycan points out that case 4 doesn't involve manipulation *by other agents*, whereas the early cases do. Perhaps it's the fact that other *agency* is interfering with our own rather than the fact that the process is causally deterministic

[85] Pereboom (2014: 79)

that's driving the intuitions of non responsibility in the early cases. If that were true then the incompatibilist wouldn't be entitled to run the argument through the cases claiming 'no relevant difference' between them. Pereboom has replied that the intuitions of non responsibility in cases 1 and 2 are just as strong when instead of manipulation by a team of scientists, the causal interference is produced by a 'spontaneously generated machine', specifically a machine with no intelligent designer.[86] A similar response is to imagine cases 1 and 2 where a lightening strike or strong electromagnetic force field cause the desired effect as has been suggested by Carolina Sartorio and Al Mele respectively. This seems right so we should conclude that it isn't the fact that *other agency is involved* that drives the non responsibility intuition. However, it should be noted that even though these considerations rule out other agency being the source of the problem, they don't rule out that in the early cases it's interference *in Plum's agency (whether that's by other agency or not)* that's driving the intuitions as opposed to the mere fact that the situations are causally determined. This last point is important given that case 4 definitely doesn't involve interference in this general sense. This point will resurface repeatedly in the discussion to follow.

Stephen Kearns sets up a dilemma for the four case argument in the following way.[87] On the first horn, Kearns' worry is that if it's the manipulation that is the source of the non responsibility intuition then this

[86] Pereboom (2001: 115)

[87] Kearns (2012: 379-89)

doesn't apply in the normal scenarios of cases 3 and 4. On the other hand if the manipulation is *not* doing this work, it must be possible to use a case that doesn't involve manipulation to elicit the non responsibility intuition. However, this would mean the manipulation examples were redundant. Patrick Todd replies to this argument as follows:

> … the proponent of the argument contends—and clearly must contend—that the manipulation is irrelevant as concerns what makes the agent unfree. She instead says that the manipulation can help us see that something does make the agent unfree. In other words, she first presents the scenario (say) to an agnostic, and asks whether the agnostic thinks that the agent is free (or responsible) in that scenario. And suppose the agnostic says 'no'. She then points out that whatever would make the agent unfree in that scenario would also make the agent unfree in a qualitatively identical scenario, except in which blind natural causes have taken the place of an intentional agent.[88]

It seems coherent that the manipulation's function in the early cases is to make salient the fact that we are causally determined and that this fact is what's really driving the intuition rather than the manipulation itself. However, we need an argument to show this or at least to make one of these positions more likely a good explanation of what's going on than the other. Todd and Pereboom need to do more than just make their reading of what's driving the intuitions coherent they need to argue it is what's actually going on.

In response to this exchange between Kearns and Todd, Pereboom sums up the dialectic as he sees it. He points out that when people first engage with the free will problem they assume that agents *are* ordinarily

[88] Todd (2013: 202)

morally responsible in the basic desert sense when they knowingly do wrong. When the idea that we might be part of a causally determined universe is presented, this doesn't challenge the assumption of responsibility for (what Pereboom calls) the 'natural compatibilist' but the thought would challenge (though not obviously defeat) the idea for the agnostic on this issue. Pereboom claims that these reactions don't adequately respond to the threat of determinism. He claims that (reiterating the intended methodology of the four case argument) that compatibilist and agnostic intuitions are more properly challenged by running a deterministic manipulation argument and then in stages (cases 1 through 4) showing that there's *no responsibility relevant difference* to be had once the manipulation is subtracted. Hence determinism is made salient as a threat. In response, this summary of the dialectic (although perhaps true) still doesn't engage with the best formulation of the challenge Kearns presents. As I said in the paragraph above, the compatibilists now claim that there *is* a candidate on the table for a responsibility relevant difference and it is precisely that the agents are being manipulated as opposed to the simple fact that they are causally determined to act as they do! I therefore reiterate that further arguments are required to establish whether (as Pereboom thinks) it's the causal determination that's driving the intuition, with the manipulation functioning merely as a way of making *that* salient, or whether it's the manipulation per say that is doing the work.

3.4 Mele's argument

Alfred Mele argues that Plum being manipulated in a 'particularly invasive way' in the first two cases is a better explanation of the non-responsibility intuitions than the fact he's causally determined.[89] Mele compares the changing intuitions to different kinds of manipulation cases (both deterministic and indeterministic) *across* the differently held positions in the debate (incompatibilist, compatibilist and agnostic). Specifically, he points out that in response to Pereboom's initial manipulation cases (cases 1 and 2), *both* compatibilists and incompatibilists have the non-responsibility intuition. Furthermore, in *indeterministic* manipulation cases it's also the case that both compatibilists and incompatibilists often share the intuition of non responsibility. The main asymmetry of intuitions arises when people consider the ordinary deterministic case (i.e. case 4) when compatibilists do *not* have the non responsibility intuition whereas incompatibilists do. Given this pattern of intuitions Mele argues that it's more plausibly the common factor of manipulation per say that's producing the non responsibility intuitions in Pereboom's initial cases than specifically causal determinism.

In response to this argument Pereboom notes that some types of manipulation of indeterministic agents *doesn't* automatically rule out us having intuitions that those agents are responsible. Specifically, an indeterministic case where manipulation enhances the strength of egotistical reasons but doesn't causally determine their bringing about a certain de-

[89] Mele (2005, 2006: 141-44; 2007)

cision (we can assume the agents have libertarian free will here) seems to leave it open that they can still (intuitively) be morally responsible. This is a problem for Mele given his claim that both sides in the debate have the intuition of non-responsibility in certain indeterministic cases. Which indeterministic cases are relevant here? Given Mele's argument it will be important to consider the exact indeterministic (libertarian) analogues of Pereboom's cases 1 and 2. What would those analogues be and what intuitions do we have about them? The indeterministic analogue of case 1 would presumably be a scenario where an egotistical reasoning process of exactly the same motivational strength is implanted just before the (now libertarian Plum) is about to make the decision. Contra Mele's argument, I think it's intuitive that libertarian Plum *is* still responsible although with some degree of diminished responsibility.[90] Why? Plausibly because the egotistical reasoning doesn't now causally determine the subsequent decision. Plum the libertarian agent as *agent-cause* determines that and hence the buck stops with him. The same goes for the relevant analogue of case 2 as far as I can see. It's important to note here that in

[90] Helen Steward has pointed out in comments that the intuition of responsibility is very plausibly diminished in this (libertarian) version of the manipulation case. I agree that responsibility can be a matter of degree in these cases depending on the strength of the desires at issue but still maintain that if the implanted/pre-programmed desires in question are not irresistible and the agent's choice isn't determined by them then the responsibility intuition, although indeed diminished, is still robust enough to constitute a problematic disanalogy with the exact deterministic analogues of these cases. Specifically, all other things equal between the deterministic and the libertarian versions of cases 1 and 2, this still allows Pereboom to claim (contra Mele) that it's the determinism and not the manipulation that's the source of the non-responsibility intuitions.

these libertarian analogues now under consideration it's the fact that Plum isn't determined and does the causing as agent-cause that's salient and not the fact that Plum 'could have done otherwise' in this scenario. Pereboom is making use of these manipulation arguments as part of his articulation of the *source* problem for compatibilism and not the leeway problem having already rejected the alternative possibilities requirement so the latter move re alternatives would not be dialectically appropriate here.[91] So the fact that both sides would agree that *some* incompatibilist manipulation cases elicit an intuition of non-responsibility won't help Mele block the four case argument here. If the relevant libertarian analogues of Pereboom's early cases are brought into focus it seems the comparison in fact serves the opposite end of helping Pereboom make it plausible that it's the manipulation *as part of a deterministic process* rather than manipulation per say that's driving the non-responsibility intuitions in these cases.

Interestingly, in his recent book, Pereboom doesn't immediately reply this way, i.e. by considering straight away the relevant indeterministic analogues of cases 1 and 2 and then arguing that country to Mele's schema there is in fact an asymmetry of intuition that comes out in his favour rather than a symmetry as Mele contends. Instead, Pereboom develops the 'dam analogy' to show that the fact that we sometimes do have intuitions that indeterministic manipulation undermines responsi-

---

[91] Thanks to Mike Otsuka for drawing my attention to this point regarding *why* we think libertarian Plum is responsible in his comments.

bility cannot automatically be used to establish that manipulation *per say* is the more plausible cause of the non-responsibility intuitions. The aim of the analogy is to show that there can be more than one type of event that can count as a token of the same general causal explanation. The very fact that the tokens are suitably different does not mean they constitute competing causal explanations so long as they both instantiate the more generic type of causal explanation at issue. The analogy is as follows:

> Imagine that a dam at one end of a reservoir would break if the reservoir were filled with more than one billion gallons of water, because the dam could not withstand the pressure that this volume of water would exert. Suppose the reservoir is in fact filled with more than one billion gallons of water, and the dam breaks. It is natural to say here: "what explains the dam's breaking is the water pressure." However, someone might object: "if the reservoir were filled with more than one billion gallons of oil, it would also have broken. So the water pressure doesn't explain the dam's breaking." To this the correct response would be: some true causal explanations set out the actual sufficient conditions for an event's occurring, and accordingly the explanation by way of the water pressure is true. But there is also an explanation of the dam's breaking common to both the water pressure and the oil pressure scenarios: liquid pressure higher than a certain level caused the dam to break. But the water-pressure explanation doesn't compete with the liquid-pressure explanation—they are explanations at different levels of generality.[92]

I agree with Pereboom that this analogy (applied to the Mele schema under consideration) makes clear that the fact that indeterministic manipulation *can* undermine responsibility is not necessarily a challenge to the claim that causally deterministic manipulation also undermines re-

[92] Pereboom (2014: 83)

sponsibility (or just causal determinism alone for that matter) *because it's deterministic,* which is the conclusion Pereboom is trying to establish. This is because, as Pereboom points out, there can be more than one way of undermining the type of control necessary for moral responsibility (just like there can be more than one way of getting the dam to burst). In summary, Pereboom suggests that when indeterministic manipulation undermines responsibility it may be because it stops the agent properly 'settling' what decision is made. In contrast, determinism can undermine the necessary control as decisions are produced by factors outside the agents control. This all seems coherent. However, it should be clear that the analogy doesn't get Pereboom to the conclusion that determinism *is* the more plausible cause of the non responsibility intuition in his cases 1 and 2. It just enables him to stop Mele running his argument to the conclusion that manipulation/interference per say is the more plausible cause of the intuitions (and not determinism). Further argument is required to show that it's one of these and not the other that's driving the intuitions in Pereboom's early cases. As I said above, I think Pereboom can establish this conclusion by examining the exact indeterministic analogues of his cases 1 and 2 and noting that we have the intuition that Plum does seem responsible in them. After the discussion of the dam analogy Pereboom does mention this type of asymmetry when he says:

> Even if, as in the kinds of cases Mele introduces, the indeterminism is only slight, this difference will be in place. Suppose in Case 1 the neuroscientists enhance the egoism of Plum's reasoning process but not quite to the degree that causally determines him to decide to kill White. Then it still may be intuitive that he is blameworthy to some

> degree, for example if the audience is imagining Plum to be an undetermined agent-cause. To get a clear non-responsibility intuition in the right sort of indeterministic case that fits Mele's description, my sense is one would need to imagine an indeterministic event-causal situation in which the probabilities conferred by the reasons on Plum's decision to kill White are very high, and not doing so very low, but he does not settle whether the decision is made. Thus my broader account cannot be undercut by pointing out that there are cases of non-responsibility in which causal determination is absent, and manipulation is present, while the manipulation results in a type of causal circumstance that precludes responsibility-relevant control. While in each type of case such control will be ruled out, the way in which it is ruled out will be different.[93]

Once again, as I see it, the vital part of this paragraph is the claim that a particular indeterministic analogue (where according to Pereboom we imagine a manipulation case with an 'undetermined *agent* cause') *is* prone to elicit an intuition of responsibility. The separate case of the indeterministic *event* causal manipulation example with the probabilities stacked very high so as to make very likely the relevant action we can concede is prone to elicit the intuition of non-responsibility but this is not the kind of case needed anyway. Firstly, it will not speak to the agent causal libertarian here and secondly, when the probabilities are high enough to prompt the non-responsibility intuition the case loses its bite as the worry that the desire is irresistible (or near enough to undermine responsibility) arises *irrespective of the issue of determinism*. This high probability event causal case is therefore not a relevantly similar indeterministic analogue (as Mele needs it to be) of the deterministic versions of Pereboom's early cases.

---

[93] Pereboom (2014: 84-5)

So far, I take the above discussion to have adequately responded to that particular part of Mele's argument which claims that there is also an intuition of non-responsibility in some *relevant* indeterministic manipulation cases. On the contrary I claim, following Pereboom's suggestion, that this simply *isn't* the case for the salient *agent causal* manipulation cases where the analogue of Plum *is* most often found intuitively responsible despite the fact that an analogously implanted desire (as in case 1) was present. However, there is more to be said here because Mele was making two important claims that taken together he took to constitute an argument that manipulation per say, rather than the fact of causal determinism, was the best explanation for the non responsibility intuition. The other part of the argument was the fact that the compatibilists *didn't* have the non responsibility intuition in the standard deterministic case 4 where no manipulation was present. Mele says the following:

> The judgment that the determinism in a deterministic manipulation case provides the best explanation of these compatibilists' 'nonresponsibility' intuitions about it is silent on the analogous indeterministic case, and it yields the prediction that these compatibilists will have 'the intuition' that Plum is not morally responsible for the killing in any straightforward deterministic story I might tell that involves no manipulation and no monkey business of any kind. Obviously, the imagined data do not warrant that prediction.[94]

I have already explained why there is no problem of 'silence on the indeterministic case' by which Mele means no *explanation* for the intuition of non responsibility in the indeterministic case. On the contrary, we standardly do have the intuition of responsibility and not the lack of it

[94] Mele (2007: 195-210)

(albeit somewhat diminished) for agent causal Plum. However, what about the fact the compatibilists have the intuition that agents are morally responsible in the standard case? Specifically, Mele claims that the incompatibilist take on things here *predicts* the intuition of non responsibility in the standard case 4. I wish to say two things in response here. Firstly, as I understand the four case argument, Pereboom was never predicting compatibilists would have any such immediate incompatibilist intuitions about case 4. Surely the idea was that working through the cases (from 1 to 4), people would be rationally compelled to adopt the status of non responsibility in case 4 given the initial non responsibility intuitions about the early cases coupled with the 'no relevant difference' claim here. So the fact that compatibilists have the intuition of responsibility in the standard case 4 isn't a problem because it goes against what Pereboom *predicts*. However, Mele can still ask whether the fact that compatibilists (and many others besides) have the intuition of responsibility in the standard case 4 is nevertheless a problem. Specifically, isn't it the case that the intuition of responsibility in the standard case 4 carries some weight here and that this must be somehow accounted for (or explained away) by Pereboom? In short, even if Pereboom gets people on board with the four case argument and the process of moving through the cases, it will be important to have not just an argument to the effect that people *should* come to hold a different intuition about the standard case 4 (this I take it exhausts the role of the Pereboom's four case argument as it stands by itself). In addition, a fully satisfactory incompatibilist response

should also have a story that *explains* why so many of us (not just compatibilists) had the intuition of responsibility in case 4 in the first place. What can be said in response to this demand?

It would seem that the incompatibilist is well placed to meet this challenge. Pereboom turns to Spinoza on this issue of the *status* of the compatibilist intuition of responsibility in the normal deterministic case 4 and says the following:

> ….However, the concern that incompatibilists have about these intuitions is that in ordinary cases we are not aware of the actual causes of our actions, and if we were, we would or should reconsider our judgments that agents are free in the sense required for moral responsibility. Spinoza observed, "men think themselves free, because they are conscious of their volitions and their appetite, and do not think, even in their dreams, of the causes by which they are disposed to wanting and willing, because they are ignorant of [those causes]" (1667/1985: 440). One serious possibility is that our choices and actions do in fact result from deterministic causal processes that trace back to factors beyond our control, while ordinary intuitions and judgments about moral responsibility do not presuppose such determinism about choice and action, and they may even presuppose that it is false. And crucially, the intuitions formed under such neutral or indeterministic suppositions might well persist even if it specified that the scenario to be assessed is deterministic (Nichols and Knobe 2007). So it stands to reason that when we are reflectively assessing manipulation arguments, the intuition that agents can be morally responsible in ordinary deterministic situations should not be accorded whatever weight we might initially assign to it. What's needed is a vehicle for making the supposition of causal determination salient in a way that effectively brings it to bear on these intuitions, judgments, and associated emotions. This is part of the point of the manipulation examples in the four-case argument. The idea is to devise scenarios in which the deterministic causes of actions are readily salient in this respect, and to show that there is no relevant difference between these causes and ordinary deterministic ones.[95]

[95] Pereboom (2014: 88)

In summary, Spinoza's point gives us pause when we consider what faith we should place in intuitions of responsibility about the standard deterministic case 4. It does not claim that intuitions of responsibility are incorrect (that would be simply question begging in the current context). Instead, it draws attention to the fact that our intuitions about the standard case should not carry the sort of weight that Mele needs them to in order to run (that part of) his argument. To take the intuitions about case 4 as informative in this context would not be to take seriously the challenge that the incompatibilist claims to be putting forward here. With this last point in place, the overall dialectic looks as follows: The intuition of non-responsibility is present in the early manipulation cases 1 and 2. There is then the no relevant difference argument from these cases to case 4 (these two claims constitute Pereboom's four case argument). At the same time many people *don't* have analogous intuitions about agent causal *in*deterministic manipulation cases. (So Mele is not entitled to claim that manipulation is the common factor across cases where we do have the non responsibility intuition, precisely because we don't have that intuition in the agent causal analogue of Pereboom's early cases). We finally have reason to place a question mark over intuitions of responsibility elicited by the standard case 4 given Spinoza's point (this casts doubt on whether we should use that intuition of responsibility (if we do have it) about case 4 in the argument as Mele does). We should be agnostic about case 4. I take it that all these taken together give Pereboom the advantage here. In particular the ball is still in the compatibilists court to

find a relevant difference across the cases. The intuitions (and crucially the status they can be accorded) seem to line up as Pereboom requires.

3.5 Bypassing and a general dialectical strategy regarding proposed compatibilist necessary conditions.

There is another response to Pereboom's four case argument that claims the manipulation at issue *bypasses* the capacities for reflective control agents have over their lives over time. The contention is that this better explains why the early manipulation cases give rise to intuitions of non responsibility rather than causal determinism. Mele and Haji have both raised this worry and Haji has developed it at length.[96] Key for Haji is that an action we're morally responsible for issues from us given an evaluative scheme we accept. That is, normative standards and deliberative principles we accept as well as the motivation to act on these principles and achieve goals based on reasoning that issues from the deliberative principles. So in the manipulation cases the crucial question is how the implanted beliefs and desires that causally determine the action stand with respect to the evaluative scheme held and endorsed. It would appear that Haji is not worried by manipulation per say a long as it's not the case that "Plum's reasoning to kill White issues from these beliefs and desires without engaging elements of Plum's authentic evaluative scheme; the reasoning bypasses these elements."[97] However, regardless

[96] Mele (2006: 166-7), Haji (2009: 166-8)

[97] Haji (2009: 167)

of whether you're a compatibilist or not, it's not plausible that we can only be morally responsible for those actions that our evaluative schemes developed over time would endorse. This point was a notorious criticism of Harry Frankfurt's hierarchical or *real self* theory of freedom. According to Frankfurt we're responsible only if our higher order desires line up with our effective first order desires (wills) and we endorse our effective desires at the second order. The 'real self' tag here signifies how the evaluative scheme and moral outlook *an agent has* is a condition on responsibility here. However, the problem for both Frankfurt's theory and Haji's requirement is that most cases of blameworthy action result from people acting at odds with their accepted evaluative scheme and normative standards due to weakness of will or because of anger etc. It's true that psychopaths sometimes endorse morally terrible schemes and then act on them but this is the exception rather than the rule. For most of us, we simply fail to live up to the goals in agency we set for ourselves. This point alone threatens to undermine Haji's criticism but in addition Pereboom counters with the immanent point that you can simply build the required evaluative endorsement Haji is after into a manipulation case. Imagine a Plum who reflectively endorses acting egotistically or immorally when it is in his self interest to a great extent. He does this every now and then on his own. Pereboom then stipulates that in case 1 Plum would not have decided to kill White *had it not been* for the scientists implanting that disposition to reason highly egotistically. Consistent with the truth of this simple counterfactual everything else Haji requires ap-

pears to be in place. The disposition clearly does not bypass Plum's endorsed evaluative scheme. The same can be analogously stipulated for case 2 and the result is the same, Haji's condition appears to be met.

The discussion of Haji's bypassing worry here highlights a general move available to Pereboom when replying to compatibilists who seek to find a morally relevant difference between the four cases by presenting a condition on responsibility they claim is violated in the early cases. There is of course no guarantee this will work but Pereboom can always try to rework the early cases so that the condition in question is now possessed by Plum. Each cases must be judged on its own terms but this potential move must always be examined as and when new (allegedly violated) conditions on responsibility are put forward as candidates for causing the intuition of non responsibility in cases 1 and 2.

3.6 McKenna's hard-line reply to manipulation cases

Michael McKenna has made the distinction between what he calls 'soft-line' and 'hard-line' replies to manipulation cases. With respect to Pereboom's four case argument, the hard-line reply takes the bullet-biting path of claiming Plum is morally responsible in all four cases. A soft-line reply instead attempts to draw a line somewhere between the different cases and claim that Plum's not responsible in the early cases but is in the later cases. So far in this chapter I have discussed compatibilist strategies that would fall under the soft-line heading but McKenna develops an argument for a hard line reply as follows. The dialectic is different from

Mele's here but the particular point McKenna leans on appears to be in tension with the Spinoza worry articulated by Pereboom above. The basic idea is that just as much as Pereboom claims he's entitled to run his argument from case 1 through to case 4, with the attendant preservation of the status of non responsibility, McKenna claims the *agnostic* would be justified in making the opposite move and (given the no morally relevant difference claim) transfer their agnosticism about case 4 back through to case 1. The aim of this point is to disarm the four case argument by pointing out that the opposite direction of travel through the cases coupled with a worry about the status of intuitions is equally justified. If this is correct then the four case argument would lose its bite without McKenna having to take a stand on whether compatibilist intuitions about the standard case 4 are justified. It would be a way of removing the advantage from the incompatibilist camp and forcing a stalemate without privileging any of the intuitions. Does this work?

Initially, this argument may seem very powerful for two reasons. Firstly, McKenna is assuming that the incompatibilist and their interlocutor now agree on the 'no relevant moral difference' premise. Secondly, unlike Mele's argument above, the compatibilist intuition of responsibility in the standard case is never appealed to. Instead, the agnostic move McKenna claims is rational about the standard case 4 looks to be just what Pereboom himself recommends given the discussion of Spinoza he appeals to in response to Mele. It therefore looks like an internal critique given what Pereboom himself says we should accept here. In reply Pere-

boom says '…if we are precise about the attitude it is rational to have about Plum in these examples, we will see that the force of this hard-line reply compatibilist response is compromised.'[98]

Pereboom clarifies what he calls the *neutral enquiring response* as opposed to the response of the *confirmed agnostic* to the standard deterministic case 4. He argues that the Spinoza considerations properly understood justify the former and not the later and that furthermore this does not result in a stalemate given that the former position *is sensitive to further clarifying considerations/changes of intuition when running through the different cases* in the four case argument. The neutral enquiring stance (NES) properly claims that it is still an open question whether the mere fact that the process in case 4 is deterministic is a reason to think Plum is not morally responsible. Now it might seem that this is just to invite a similar argument for stalemate couched in terms of NES rather than whatever species of agnosticism McKenna had in mind. However, anticipating the reiteration of the argument in terms of the NES, Pereboom says:

> …. this suggestion does not take into account that adducing an analogy for which one's intuitions are clearer might itself count as the relevant sort of clarifying consideration. If the neutral inquiring response to an ordinary causally determined agent were initially epistemically rational, it might then be that an analogous manipulation case functions as a clarifying consideration that makes rational the belief that the ordinary causally determined agent is not morally re-

[98] Pereboom (2014: 91)

> sponsible. The confirmed agnostic rules out this possibility, but not the neutral inquirer.[99]

The basic idea is that we should start off with NES upon consideration of case 4 and then move from case 4 back towards case 1. Once our intuitions change to clearly indicate that Plum 1 and 2 are not responsible, we are then entitled to reason *back* through to the standard case 4 via the no-relevant-difference premise and conclude that determinism undermines responsibility. This is not an ad hoc move given the initial comparative function of the four case argument in the first place. It won't work if someone never gets intuitions of non-responsibility in cases 1 and 2, but it is widely accepted now that many confirmed compatibilists/agnostics do have such intuitions and are therefore in need of a principled way of explaining them away. In summary, the most epistemically rational position (regardless of your stated view on the free will problem) is to take up the NES with respect to the standard deterministic case and then move between the cases until 'clarifying considerations' become available. If they do, i.e. in the form of a firm intuition of non responsibility in the standard case then we're in a position to pass judgement on the standard deterministic case having run back and forth through the cases as described. Pereboom is thus able to maintain the dialectical advantage here and the stalemate that McKenna seeks is avoided.

---

[99] Pereboom (2014: 94)

In response McKenna agrees with Pereboom that the initial stance to the standard case should be open to further considerations as the NES is. However, he counters that the intuitions about the standard case 4 should carry some weight when considered alongside the intuitions about the manipulation cases 1 and 2. The idea being that we should place more faith in our intuitions about ordinary cases than we do about the contrived manipulation cases. In response to this move it seems correct to reiterate the point from Spinoza which aims to bring to bear considerations that challenge the status of our intuitions here. Again, Spinoza's worry is that we should not trust our ordinary judgements because we don't ordinarily conceive our actions as part of a deterministic process and what's more, we may well be operating with a tacit assumption of libertarian free will and moral responsibility. The fact that it's *stipulated* that the ordinary case 4 is deterministic is still no guarantee that our intuitions will respond adequately given the deeply ingrained nature of the way we conceive of our place in nature. I concur that this response rebuts McKenna's claim that intuitions about the ordinary deterministic cases are more reliable but what about the other claim that we should be wary of our intuitions about artificial cases? To this worry Pereboom replies that the artificially of the manipulation cases is required to make salient the fact of deterministic causation, given that (as is made clear by the first point) this supposition is 'readily suppressed' in our thinking about the ordinary deterministic case. Talking about the manipulation examples, Dana Nelkin goes on to suggest that, "one might

argue that their unrealistic quality helps ensure that we are focused on the stipulated features, and that we aren't implicitly but unconsciously relying on background assumptions that we bring to ordinary life. In this way, the intuitions are arguably *more* reliable than the real life ones."[100] Perhaps it's exactly the opposite way round here once we have properly taken onboard Spinoza's worry! Plausibly then, our intuitions about manipulation cases are more epistemically respectable than those about the standard case. In conclusion, it seems Pereboom still has the advantage here.

At this point, it might be tempting for the compatibilist to take a different tack when trying to formulate a hard line reply. Specifically, couldn't a compatibilist feasibly have the intuition that Professor Plum in cases 1 and 2 *is* straightforwardly morally responsible in the normal way? It will be impossible to deploy the four case argument against such a person in order to change their mind. You wouldn't get any leverage against them at all. This might seem like a implausible way to go but such a reply has some historical precedent. Pereboom considers the New England Congregationalist theologian Nathaniel Emmons (1745-1840) and his sermon on the bible story of Pharaoh and Moses. The main idea is that even though God directly hardens Pharaoh's heart with respect to the treatment of Moses and the Israelites in Exodus (8-14), (this is the part which is meant to be analogous to the manipulation cases i.e. Plum in case 1), to forgive Pharaoh and not hold him morally responsible

[100] Nelkin (2012)

would be a mistake. The sermon Pereboom draws our attention to runs as follows:

> It is often thought and said that nothing more was necessary on God's part, in order to fit Pharaoh for destruction, than barely to leave him to himself. But God knew that no external means and motives would be sufficient of themselves to form his moral character. He was determined, therefore, to operate on his heart itself, and cause him to put forth certain evil exercises in the view of certain external motives. When Moses called upon him to let the people go, God stood by him and moved him to refuse. When Moses interceded for him and procured him respite, God stood by him and moved him to exult in his obstinacy. When the people departed from his kingdom, God stood by him and moved him to pursue after them with increased malice and revenge. And what God did on such particular occasions, he did at all times. He continually hardened his heart, and governed all the exercises of his mind, from the day of his birth to the day of his death. This was absolutely necessary to prepare him for his final state. All other methods, without this, would have failed of fitting him for his destruction … Pharaoh was a reprobate. God determined him from eternity to make him finally miserable. This determination he eventually carried into effect. He brought him into being, formed him a rational and accountable creature, tried him with mercies and judgments, hardened his heart under both, caused him to fill up the measure of his iniquity, and finally cut him off by an act of justice.[101]

Pereboom points out that many other Calvinists might share these hard line intuitions as well. What are we to make of this? It is clearly a local manipulation case with God and Pharaoh, a situation analogous to Plum in case 1. To start with I think it's worth pointing out that a great many compatibilists, as well as agnostics do not take this approach, presumably given they don't have the intuition that Pharaoh and Plum 1&2 are responsible. As for those that genuinely do have these intuitions, one

[101] Emmons, v. 2., (1860), 327, 330/(1987), 391–2; 395.

argument that can be brought to bear on them following empirical work done by Shaun Nichols and Joshua Knobe, is that these cases involve grossly morally bad actions (regardless of whether the agent in question, i.e. Pharaoh or Plum is morally responsible) and that this fact has been shown to sway intuitions *towards* holding people responsible compared to contexts where subjects otherwise exhibit intuitive reactions to cases indicative of incompatibilism.[102] This is hardly a knockdown argument here but it is a worry about the reliability of intuitive responses in these extreme cases, extreme in terms of the badness of the deeds. Pereboom, discussing Nichols and Knobe, says, "…An explanation they consider is that in the high affect cases, the rational response is suppressed or precluded by emotion, here perhaps indignation or vengefulness. But one might argue instead that in the low affect cases, inappropriate sympathy suppresses the rational response, while in the high affect cases, inappropriate sympathy is more likely to be balanced off by indignation, allowing the appropriate response to emerge."[103] I personally feel that the original suggestion of emotion clouding otherwise incompatibilist intuition is more plausible than the compatibilist move suggested by Pereboom - why would 'inappropriate sympathy' emerge in the first place - *because* the agents in question are understood to be causally determined? However, I concede I don't have an adequate response to this position apart from to point out it's rarely taken in the contemporary debate. It's very

[102] Nichols and Knobe (2007)

[103] Pereboom (2014: 97)

difficult to deal with someone who honestly reports these hard-line intuitions of responsibility in the manipulation cases but most people don't and it's fair to say that many compatibilists do see the manipulation arguments and the intuitions they give rise to as serious problems that need to be responded to in defence of their position.

3.7 A worry about asymmetry — Do Pereboom's manipulation cases deliver the same result for morally good action? If not, is this a problem?

In the literature both Susan Wolf and Dana Nelkin defend the position that moral responsibility requires the ability to do the right thing for the right reasons. The consequences of this are asymmetrical in that blameworthy action requires the existence of an alternative (the right path *not* taken) but praiseworthy action doesn't. In contrast to this, Pereboom's position is that moral responsibility (in the basic desert sense) for *both* blameworthy and praiseworthy actions is incompatible with causal determinism. Given this, shouldn't we expect the manipulation arguments to deliver the same results for praiseworthy action as for blameworthy action? It looks however, like some of the instances of manipulated good behaviour tend to elicit intuitions of responsibility, in contrast to their exact analogues for blameworthy action. For example, Nelkin says;

> … it would be instructive to vary the example so that Plum acts well. Suppose for example that we replace case 2 with this one. The neuroscientists have created Plum so that he fully understands and recognizes good reasons for acting. He is moved by people in distress and

> desires to help them so as to relieve their suffering. Suppose that one day he finds himself in a situation in which he can help a child only at great risk to himself. He thinks about the relevant considerations and decides that all things considered, helping is the right thing to do and resolves to do it. The intuitive force of this example seems to me to swing the other way. At the very least, it is much less appealing to say that Plum is not responsible. If indeed there is an asymmetry in our reactions to the cases, it suggests an explanation other than that we find people not to be responsible when their choices are determined by causes beyond their control. What capacities one has when one acts is what is essential here. This line of response both questions whether cases 1–3 serve as independent counterexamples to the rational abilities view, and also points to an alternative explanation for our intuitions that allows us to resist the general conclusion against compatibilism.[104]

Another example along these lines is McKenna's discussion of a woman who is causally determined and also (see the discussion) manipulated into making the hard choices to live a morally worthy and fulfilling life in very difficult circumstances. As with Nelkin's take on her version of Plum 2, McKenna says of the striving woman in his example that intuitively she's morally responsible for living her life and choosing as she does and that this intuition is likely to be widespread in responses to the case.

I think Nelkin and McKenna are right that the manipulation cases for good actions tend to produce intuitions of responsibility more often than their counterparts involving bad actions. Pereboom himself concedes that it seems intuitively more appropriate to praise in the 'good action' manipulation cases than it does to blame in the 'bad action' ones but nevertheless contends that this doesn't amount to showing that moral

[104] Nelkin (2011)

praise is deserved *in the basic desert sense* in these cases. In defence of this claim he explains that many characteristics people have typically elicit praise when it's at the same time quite clear that the people with those characteristics don't fundamentally (in the basic desert sense) deserve praise for them. Examples include, good looks, athletic ability and intelligence. The key point here is that these characteristics don't even involve agential activity so it's not surprising that people tend to praise Nelkin's courageous Plum or McKenna's determined woman *if they are already happy to praise for (involuntary) desirable characteristics in general*. To this Pereboom adds the further point that there is an asymmetry in the appropriateness of holding people blameworthy and praiseworthy based on the fact that no great harm is done if people are praised for things when they don't necessarily fundamentally deserve to be (in fact this may well be what we should do here regardless) but that seems much less true of blame given that underserved blaming seems wrong from the get go. The explanation being that in the latter case harm is done to a person blamed when they don't deserve it.[105] Given these considerations Pereboom concludes that the (acknowledged) asymmetrical reaction to manipulation cases involving good and bad actions doesn't undermine his claim that moral responsibility in the basic desert entailing sense is incompatible with causal determinism for all agency, both good and bad.

[105] For further discussion of this asymmetry see Pereboom (2001: 139-41)

Chapter 4

4. Revisionism about moral responsibility

4.1 Diagnostic incompatibilism, prescriptive compatibilism

In this chapter, taking my lead from the recent work of Manuel Vargas, I will explore the options for revisionism about moral responsibility.[106] By revisionism I mean the idea that our concepts can be modified as well as simply maintained as they are or rejected altogether. They should be revised when, while part of their content needs to be rejected, sufficient content remains such that the concept is still justified in being applied (post revision). The extent to which we are justified in revising concepts will be a matter of degree. The possibility of conceptual revision promises to help in the apparent stalemate between compatibilists and incompatibilists in the free will problem because a revisionist position has the resources to accept (some of the) central claims made by both parties in the debate. More specifically, revisionism can concede the semantic point to the incompatibilist (that natural language uses of responsibility presuppose the falsity of determinism) while at the same time maintaining that sufficient inferential import of the concepts at issue remains intact for us to be justified with continuing to deploy them.

The familiar distinction between compatibilism and incompatibilism is often applied without it having been made clear exactly *what* is sup-

[106] Vargas (2013)

posed to be compatible or incompatible with determinism. Is the compatibilist claiming that our natural language concepts of free will and moral responsibility are consistent with the truth of determinism, or instead that we *should* be compatibilists about freedom and responsibility, even if there are incompatibilist commitments in our folk conceptual scheme (i.e. even if what it normally means to say someone is morally responsible assumes the falsity of determinism). Both views are fairly called compatibilist but they are different positions and you can hold one while rejecting the other. With respect to the concept of and characteristic practices associated with moral responsibility, a *diagnostic* view is simply your take on whether the concepts as they stand in natural language and practice do or do not presuppose the truth of determinism. A *prescriptive* view is a position on whether we *should* be deploying a concept of responsibility (perhaps suitably modified, though not necessarily) whatever diagnostic position is adopted.

The purpose of this chapter is to explore the options for developing the second of these views. Working on the assumption that there are recalcitrant incompatibilist semantics in our concept of moral responsibility (diagnostic incompatibilism), many have taken the view that if determinism were true we would have to deny there was any responsibility in the world. However, this is too quick because we can always be revisionists about our concepts, if sufficient justifications for revision exist. Sometimes when a concept's commitments are found to be unrealisable we are duty bound to stop deploying it. There are also situations where suffi-

cient referential or connotative content survives (albeit after *some* of the content has been found wanting) to be justified in continuing to deploy that concept. I will evaluate what form revisionism for the concept of moral responsibility might take and our justifications for persisting with the concept.[107]

Working on the assumption of diagnostic incompatibilism, i.e. assuming that one of the conditions on moral responsibility as we currently understand it requires the falsity of determinism, what justifications might there be for prescriptive compatibilism? In what follows I will explore two main options, firstly, the idea that even if folk incompatibilism is true, it might nevertheless be true that our concept of moral responsibility does a good job at organising our distinctively moral practice. This would be a reason to revise and then retain it in the face of diagnostic incompatibilism. Secondly, what I shall call the 'freestanding normative argument' or FNA, claims that we need to preserve the idea and practice (albeit suitably revised) of moral responsibility because it plays a vital role in the realisation of what we value in human community and social

[107] Interestingly, there is space to be both a diagnostic incompatibilist/prescriptive compatibilist as well as a diagnostic compatibilist/prescriptive incompatibilist! The second position is weird but coherent. Imagine a person who grew up in a culture where the folk concept of responsibility was compatibilist, yet who came to believe (perhaps due to theological commitments about sin and punishment) that we should in fact be deploying an incompatibilist notion of responsibility! The more familiar positions are diagnostic incompatibilism/libertarian commitment, diagnostic incompatibilism/responsibility scepticism and straightforward diagnostic compatibilism (i.e., Harry Frankfurt, G.E. Moore's conditional analysis compatibilism). Manuel Vargas is responsible for organising the debate around these distinctions in his *Building Better Beings* (2013).

flourishing. These two justifications are not the same as it might be that the former line of argument alone could not supply adequate justification where the second might, i.e. even if our revised concept of responsibility (shorn of incompatibilist content) did not manage to systematically order our *existing* judgements in an adequate way, we may still be justified in adopting a revised concept if it was necessary to achieve an independently motivated human flourishing. In what follows I hope to make clear how these arguments differ and also the extent to which they can be used together in a revisionist theory.

It is useful to look at instances of revisionism in other areas before looking at whether analogous changes can be made to the concept of moral responsibility. Manuel Vargas claims:

> Nearly any significant concept – physical, moral, or otherwise – that has a long enough history is unlikely to survive unrevised in the face of growing knowledge about the world. Given that the notion of simultaneity proper to physics, our conception of water's essence, our moral notions of what constitute virtues, and our conception of marriage have all been subject to revisions in various ways, it is not obvious why we must suppose that our concept of responsibility is immune to revision. Indeed, given the particular sociocultural history of the idea of responsibility, and in particular, the role it played in Christian theology and pre-scientific conceptions of the self, it seems unduly optimistic to suppose that this particular culturally inherited concept will have come down to us in a form that is smoothly compatible with a contemporary scientific view of the world.[108]

[108] Vargas (2013: 76)

4.2 What is revisionism?

Vargas outlines various ways conceptual revision might occur. He illustrates these differences by considering the traditional problem of meta-ethics. Originally many people held a divine command theory of morality, which claimed that the content of morality is (perhaps also necessarily) determined by the will of God. As people started to reject theism the status of ethical and moral claims became the subject of debate. If you believed that what it means to say something was immoral was that God had said so, would this mean that you *must* become a moral nihilist/sceptic once you also became an atheist/agnostic? It is now widely agreed that this is not the only option and that what needs to be rejected is rather the thought that ethics requires God. People instead came to believe that the *pattern* of judgements and practices typical of morality can find justification in practical reason or as a means to the end of achieving desirable functioning and community. Deontological, contractualist and consequentialist justifications of various types have been put forward to this end. The key point is that we are still talking about *ethics*, not some hollow analogue of ethics instead, call it *ethics** for example. Specifically, we are justified in saying we're talking about the same concept because enough of the inferential role of the concept is preserved. What enough comes to will be a matter of debate and borderline in some cases but it should be clear enough in many cases as to when we do have adequate

justification and when we don't. I will expand on this point about inferential role preservation below.

We might understand revisionism about the status of ethics in two ways. Vargas makes the distinction between *connotational* and *denotational* revisionism. In the former case you would be giving up on content that doesn't play a reference fixing role with respect to the concept in question. Returning to the example of revisionism in meta-ethics, the semantics of moral claims would be different but the referents will be the same. The set of wrong acts is the same but the terms used to refer to them would have (in part) a different sense, based on the meta-ethics adopted. To reiterate the point made at the end of the last paragraph, we would still be talking about *ethics* and not something else instead, because inferential role would be largely preserved. A second possibility is that some of the content that fixes reference is itself given up. This, Vargas concedes is the trickier option to make coherent as it might just seem obvious that nihilism follows from denotational revision. After all, wouldn't we be referring to something other than we used to be? However, there is still space to call this type of denotational revision a revision of our existing moral concepts and not the adoption of new ones. Of course (in one sense) it is the adoption of new concepts, if you are intent on calling any concept that has undergone change a new concept. However, it's also true that after both connotational and denotational revision, the moral concepts in question can still play largely the same inferential roles with respect to how they organise our beliefs and practices – how they func-

tion with respect to our broad moral commitment. Vargas concedes that even though denotational revision in one sense, 'changes the topic', it still respects what he calls 'the work of the concept'. Vargas defines 'work of the concept' as "the primary inferential roles figuring in those forms of life connected to the uses of the term, namely the import of morality."[109] It is important to make clear that what makes denotational revision possible is that there be some other property which is in all the (or perhaps most of the) places that the old referents of our concepts were supposed to be. Furthermore this property licences the inferences, beliefs and social practices that are characteristic of our moral lives.

4.3 Examples of concept revision

Some of the best known examples of concept revision concern natural kind terms. We were always referring to the same stuff when we talked about water but we now know that we used to be wrong about water's essence. Water is H_2O and not whatever it was taken to be before. Concerning the inferential role of the term water, after it was discovered to be H_20 most of the old inferences we made using the concept were preserved, such as; people and plants will die without water, you can't walk on water, water is transparent etc. Some where not though, specifically whatever the old theory said about water's essence, inferences about that had to be rejected in favour of different ones based on the new science. The key point is that enough of the inferential role (in fact the vast major-

[109] See Vargas (2013: 88–9)

ity) was preserved such that we were justified in continuing to deploy the concept of water in light of the revised views on its essence. What about concepts that are not natural kind terms? I will consider some controversial examples and non-controversial examples in order to better see what the prospects for revisionism about responsibility look like.

4.4 Controversial revision

The concept of marriage has undergone significant revision over time. Originally, it was (partly) taken to mean a property relationship between man and wife, the husband coming to own the wife after the bride's father gives her away. Thankfully, most people have revised this sense of marriage and instead shifted to a concept implying a 'legally sanctioned and privileged relationship'. However, what has been a lot more controversial in recent decades is whether this later rendering is also appropriately applied to same sex couples, i.e. can gay couples get married? I wanted to bring the example of the gay marriage debate into the picture because I think it is very useful to compare the type of disagreement in that debate with the type found in whether we could or should revise our concept of moral responsibility to a thoroughgoing compatibilist sense. Specifically, when a conservative Christian says of legally sanctioned gay partnerships, 'but that just isn't *marriage*', what are the parallels between that stance and the reasons for holding it (and the reason's it may well be unjustified) and the incompatibilist who says 'but that

wouldn't be *real* responsibility' to the prescriptive compatibilists suggested conceptual revision. I will explore these (dis)analogies below.

4.5 Non-controversial revision

Scientific concepts are perhaps the best place to look for examples of uncontroversially revised concepts. Vargas gives the example of the notion of simultaneity proper to physics. After Einstein's theories of relativity, that concept had to be revised accordingly. As with the examples of natural kinds above, certain inferences (notably those directly at odds with relativity) had to be given up on. Even so, within the new scientific framework, the concept was still able to do a lot of work very similar to what it had done before (or mostly the same work – *suitably qualified* (if you prefer)) and was kept on. People didn't say 'nothing's really simultaneous anymore'. They understood that that concept meant something slightly different now, but that at the same time enough of the inferential role could be preserved in order to recognise it as the same (though revised) concept.

4.6 Revisionism and concept rejection

It is important to note at this point that the nature of revisionism described above does not guarantee we will always be able to revise our concepts (or that we should do so) in the face of new information. It is always possible that a concept may be found to be so wanting that it is

better to discard it altogether. An example from science to illustrate this point is the concept of phlogiston. Vargas discusses Phlogiston, which was thought to be a substance present in matter that made combustion possible and which was released when matter was burnt. As more and more experimental data became available it became harder to continue to underwrite the concept. Some metals were found to *gain* weight when burned etc. The structure of the explanation/model that phlogiston featured in was shown to be false. In the end the concept was discarded from science and what it was supposed to explain was better explained as the reaction with oxygen. It might be said at this point that at a suitably abstract level of description we might have had reason to revise and continue to use the concept. If the concept was meant to pick out (the *key*? ingredient) that made combustion possible then perhaps we might have re-fixed the referent to oxygen and some different process of chemistry? Why didn't we do this instead of dropping it altogether? The answer is surely that phlogiston didn't pick out enough (in this case anything specific at all) in the world. It was therefore more sensible to signify this fact by deploying new concepts altogether and dropping the label. What did do the work in explaining combustion didn't look anything like the purported referents of the term phlogiston. This is consistent with the fact that there were *some* things that did what phlogiston was trying to explain. The point is that the new science rendered false most of the inferences that used to be made with the concept phlogiston; the inferential role of the concept was not worth saving. There is a spectrum of

possibility here and no clear cut answer in every case as to when we should persist with (revised) concepts and when we should reject them. It's just important to note that nothing about the methodology of revision described above rules out discarding concepts.

4.7 A worry about claims of essentialism blocking revision

Given the above examples of natural kind and scientific concepts that have been revised as well as phlogiston being rejected, it should be roughly clear as to the overall methodology of revision. If *enough* of the inferential role can be preserved then the concept gets to stay. The inferential roles of the concept pre and post revision will not be the same but if they are sufficiently similar it makes sense to stick with the concept given the work it does for us in organising our ideas and practice. The fact that it's vague as to how much of the inferential role needs to be preserved in various cases is not a problem for the claim that this is how revision proceeds. Each case must be judged individually.

Having set out why revision occurs, there is a worry that the incompatibilist can nevertheless take a position which counts against the possibility of revising the concept of moral responsibility. Without yet having argued that some suitable compatibilist notion of moral responsibility would preserve the important majority of the existing inferential role of the concept, what if the incompatibilist claimed that the incompatibilist content was *essential* to the content of the concept? Going back to the example of gay marriage above, this would be analogous to the fun-

damentalist Christian saying it was essential that marriage was between a man and woman, *even if the rest of the inferential role of the concept could be preserved* for gay marriage. In response to this, I think the compatibilist is entitled to claim that the dispute would be merely verbal in that case. The important point is that the majority of the inferential role has been preserved and we are therefore justified in persisting with the concept that anchors this remaining inferential role – *call it what you like the compatibilist might say, but it makes sense to carry on calling it responsibility given how it continues to fit in with the rest of our conceptual scheme!* Of course, if the incompatibilist has an independent argument as to why the remaining inferential role the compatibilist is claiming persists post revision is still implicitly assuming incompatibilist content then they would have an argument for the incompatibilist content being essential to the concept, but without that, the insistence on a particular component of the old concept being essential would amount to a verbal dispute. I say that given the methodology of revisionism under consideration here. The incompatibilist has to either give that argument or present an argument about why the general method of revision employed here is unsound. Failing either of those the dispute is verbal.

4.8 Related issues with reference

It is also worth noting that the possibility of revising concepts does not hinge on the outcome of the internalist/externalist debates in the theory of meaning. On internalist theories of meaning where reference is set by

sense or beliefs about what responsibility is, we might change some of those beliefs while at the same time retaining enough beliefs about the concept such that reference is preserved, i.e. in such a way that what we then go on to refer to are the same actions, practices and patterns in the world. On externalist models of reference, reference is determined by more than just our sense of the concept, for example ostention and causal links independently of our beliefs. It is easier to see how revision is possible for externalists than internalists about meaning. If the inferential role or work of the concept *just is* what secures the reference of the term then this will arguably be robust in the face of us dropping some beliefs about libertarian powers we associated with the concept of responsibility assuming the compatibilist remainder is sufficient. Either way we can revise our concepts.

To summarise these distinctions and return to the topic of free will and responsibility, the connotative revisionist thinks that some of our folk beliefs about free will and responsibility need to be revised but that doing this would not interfere with the reference of these concepts. On the other hand, the denotational revisionist does think we need to change the reference of the terms but that we can re-fix the reference of these concepts in ways that preserves the core inferential roles, practices and beliefs associated with moral responsibility. Vargas uses the expression 'earning the survival of the term'. How do we earn the survival of the term 'moral responsibility'?

4.9 Arguments for revisionism about free will and responsibility

If we made the assumption that one of the conditions on responsibility was incompatibilist, it would still be the case that this condition was *just one* among a set of other necessary and jointly sufficient conditions for responsibility. Incompatibilist theories need these other conditions fully fleshed out as much as their compatibilist rivals. The other conditions here are familiar, being epistemic, concerning the degree of moral understanding, the capacity for intentional agency and also the reasons responsiveness of agents. One key question is therefore whether the other conditions taken together can be used to ground revisionism with respect to the role of the concept of moral responsibility. I will explore two options below. Firstly, we need to see whether the remaining body of conditions can still manage to refer to and explain the characteristic body of acts, judgements and practices when the incompatibilist control condition is removed.[110] Secondly, we need to see whether that incompatibilist control condition permits of a compatibilist reading/analogue which can be adopted instead and that then would (along with the other conditions) allow sufficient inferential role to be preserved.

Regarding the first of the above options, the remaining conditions are compatible with determinism and together they do seem to carve out much of the use and inferential role of the concept of responsibility re-

[110] I say this consistent with the thought that there are other control conditions, i.e. compatibilist ones.

gardless of whether an additional incompatibilist condition is realised or not. For example when we look to explain why we moderated our reactive attitudes in one case and not another, the difference might be due to a contrast between wilful harm infliction and accidental harm (intentional action). Perhaps another difference in the degree of blame we feel is appropriate to agents could be due to the fact that one was an adult and one only a child who had not yet developed a full moral sense (epistemic). Excusing conditions that don't make reference to alternative possibilities are to the point here. The familiar theories of the reactive attitudes developed by P. F. Strawson and R. J. Wallace are available to this end and can work to secure that portion of the inferential role without any alternate possibility condition alongside them. If the core arguments of this thesis are correct then Wallace and Strawson didn't manage to establish compatibilism by the traditional diagnostic route. However, working in a prescriptive vein as a revisionist it is perfectly acceptable to embrace their accounts of the reactive attitudes. If those accounts capture much of the inferential role of moral responsibility without any leeway conditions then that may well be enough to secure the 'survival of the term'.

Alternatively, given the second option above, we can embrace a 'could have done otherwise' condition alongside the others but read it in a compatibilist way, like List does. It might be thought that a theory of responsibility that offers no treatment of 'could have done otherwise' is too radical a departure even for a revisionist project. The worry would be

that much of the inferential role of the concept would be lost without some rendering of the alternatives requirement here. This problem won't arise if you embrace something like Wallace's view but given that many people don't, I need to say something about the scope for a revisionism that does build in a (compatibilist friendly) alternative possibility condition. Once again, what is required here is that the set of conditions taken together are s*ufficient for the preservation of the majority of the inferential role*. In other words, the compatibilist reading is enough to allow the concept to do what it needs to do. It can be conceded to the incompatibilist that a metaphysical reading of possibility might also be required by the folk semantics, but the claim is that the absence of this from the revised concept still allows the concept to organise our moral practice. List's account of modal semantics looks to be an especially good candidate here given the problems with traditional conditional compatibilism and new dispositionalism.

4.10 The argument for the preservation of (most of) the inferential role of moral responsibility on a compatibilist reading of 'could have done otherwise'

Take any familiar set of incompatibilist conditions taken to be necessary and sufficient for moral responsibility. We can read the notorious modal conditions that require the agent 'could have done otherwise' in an agentive rather than a metaphysical sense (List would say microphysical). We can also read the source hood requirement (however this is fleshed out)

in a compatibilist rather than an incompatibilist way. Conceding whatever the incompatibilist insists on regarding the natural language semantics of agency, source hood and the ability to do otherwise, once we adopt the compatibilist reading of these conditions the claim is that the revised concept is still able to function in broadly the same way as the existing one. I mean by this that the revised concept has the resources to explain the distinctive pattern of intuitions we have about cases where people are/aren't responsible.[111] Specifically, those kinds of cases where we are worried that an agent is not responsible because they couldn't have done otherwise will depend on (roughly) whether or not they had a reason to do otherwise that (*for all they knew, or should have known)* they could have acted on. If they did and yet they still did the wrong thing they may be responsible, if not, not. Worries about source hood should be understood in terms of the right kind of causal chain from intention to action that doesn't involve manipulation etc. Why should we think the agentive reading of the 'could have done otherwise' condition does what I claim it can? The key question is about cases where we do hold people responsible; is there in the agentive sense something they could (for all they knew) have done instead or should have known about? It seems highly plausible that in almost all cases there will be something we expected the

[111] This sentence may sound like a straightforward refutation of the revisionist position. If such a revised concept is able to 'explain the pattern of our existing intuitions' then why isn't it also to the same extent a good argument for *diagnostic* compatibilism? What I mean here is the claim that the pattern of intuitions will be recognisably the same across the cases but absent the *basic desert sense* of moral responsibility. The rest of the variations and nuances can be preserved with basic desert removed from all responsibility attributions.

agent to have (for all they knew) tried to do instead. If there was no way of them climbing over the fence to save the drowning child then they are not to blame but if they (for all they knew and consistent with their general agentive ability) could have climbed over and yet didn't try to, then they are to be held responsible and so on for a variety of circumstances. I think that for most cases where the possibility of doing otherwise is salient, the agentive reading of 'could have done otherwise' will allow us to get the right answer about the case and hence the concept of responsibility to maintain its core inferential role. *The main reason for this is because the metaphysical reading of 'could' that the incompatibilist claims is necessary normally runs together with the agentive sense of 'could' that the compatibilist thinks is alone sufficient*. What I mean by this is that when the incompatibilist says 'they could/couldn't do otherwise' they normally invoke both the senses without making the distinction, i.e. the incompatibilist thinks that in all those cases where there *was* something the agent could (for all they knew) have done otherwise, the agent is responsible (and of course they also assume that metaphysical possibility is present as well, the latter is the supervenience base for agentive possibility if you like). Hence we can just run the agentive sense alone where responsibility is intuitively present. Where the agentive possibility *isn't* there (barring cases of indirect (derivative) responsibility and tracing) the agent won't have been responsible, whether or not there was a bare metaphysical possibility of acting otherwise and this is something incompatibilist and compatibilist can agree on.

In summary then, for any case we might consider, if it's salient that we establish whether the agent could have done otherwise, we can use the agentive sense of could alone and get the right answer about the case. The justification for this claim is that this sense of ability is always there in those cases where basic desert responsibility was traditionally assigned because the agent could have done otherwise in a sense presumed to run hand in hand with metaphysical possibility itself yet distinct from it. If they couldn't do otherwise in the agentive sense, then both sides of the debate will agree that they can't be responsible anyway, i.e if the agentive possibility *isn't* there then that is sufficient to rule out responsibility whether or not there was a metaphysical possibility as well. It's important to remember that in a revisionist context I don't need to argue that this agentive reading alone is sufficient to deploy the old concept (I have argued at length it isn't), just that if we use the concept this way we can preserve the majority of the inferential role of the old concept minus the basic desert ascriptions. This much seems plausibly true.

At this point the incompatibilist critic might well throw up their hands. They might grant all I have said in the above two paragraphs but say that it is beside the point. No one is *really* responsible they might say. Even *if* the majority of our moral judgement and practice is still tracked and referred to by the revised concept no one is *really* the author of their actions, we're all determined to do what we do for good or bad etc, hence what's the sense in holding people responsible? This appeal to intuition

is not just rhetorical and I think every intellectually honest compatibilist should be candid about the hold such claims have on us even after working on this problem for a good while. It is a serious challenge and requires the revisionist to explain the rational for continuing with the concept, i.e. crucially it's a demand for justification *independently of the fact we might have demonstrated we can preserve a sufficient portion of the inferential role*, 'the work of the concept' as Vargas would call it, without the incompatibilist reading of 'could have done otherwise'. What's really being expressed here is something like a request for justification of moral practice *in general*. Presumably because without incompatibilist metaphysics the suspicion is that moral/value assignments don't make any sense. Another way of making the point is to say that even granting we have preserved 'the work of the concept' there is no deep justification for doing 'that kind of work' anymore. I will try to answer this question in detail in the next section.

The worry in the paragraph above is linked to the worry about claims of essentialism for the incompatibilist conditions that I mentioned in a previous section. In that section I classified the (without argument) insistence on the incompatibilist condition being essential as causing a merely verbal dispute. The reason I gave for this was that if (without the incompatibilist condition) the majority of inferential role can be preserved then all we would be arguing over is what to call the revised concept that was doing the majority of the work that the old concept did. This would be a verbal rather than a substantive disagreement. I stand by this but it's

now possible to read the claim of essentialism in light of this 'throwing up of hands worry' above. Perhaps the thought (and again the incompatibilist will have to argue for it) is that the claim of essentialism was an expression of the idea that libertarian metaphysics are required for us to make sense moral practice in general. It is an interesting question as to where the burden of explanation lies here. We are not arguing about whether our revised concept of responsibility requires incompatibilism, by stipulation it doesn't, the point is that the incompatibilists want to be told the general rationale, all things considered, so to speak, for moving forward with our revised concept. We need an argument that gets a handle on why moral practice and specifically the practice of holding people morally responsible (in the revised sense) is justified in general.

4.11 The Argument for Prescriptive Compatibilism

I now turn to answering the worry outlined at the end of the last section. Could we argue that we should be compatibilists because we *need* a concept of moral responsibility? The justification for using our revised concept would therefore be because we had to. This necessity would have to be of a kind that spoke to the worry that there's no real point if it's 'all in the cards', no point if determinism is true etc. Granting that we can preserve the majority inferential role of the concept, why should we continue with it? This is not an easy question to answer satisfactorily but it also isn't a mystery as to what kinds of justification should be forthcoming. Presumably we couldn't do without the concept of moral responsibility

given some plausible foundational assumptions about what's constitutive of the good and flourishing human life and the kind of practices, communication and interaction that's required to maintain it? This isn't (by invoking value claims) to beg the question against the general worry, as it seems reasonable to suggest there are sources of value that don't depend on libertarian metaphysics anyway. Basically, without the responsibility system and its concepts and practices all hell would (in the end) break loose. Nobody wants that, regardless of how much they thought that libertarian metaphysics was the 'font of value'. It is possible for someone to bite the bullet on this point and insist we should stop the whole practice of responsibility but I don't see this possibility as dialectically that problematic for a revisionary compatibilist here. Presumably most people will get on board with this argument (*if* we do in fact need the responsibility system, given our desires to live in a civilised world).

The key idea is that we need the responsibility system to regulate our behaviour. It is a way of communicating when people have done good or bad things, in the hope that that they will realise the significance of their acts and behave differently in the future. It's important to do this as without it we would descend into a barbaric world, unpredictable and dangerous where humans couldn't flourish and develop higher forms of culture. The responsibility system with its characteristic attributions and feedback is a moral conversation we continually have with each other to the end of living more harmonious better lives for us and those around us. This idea might look like it's subject to P. F. Strawson's classic worry

outlined in Freedom and Resentment about this kind of move. The 'efficacy' of regulating our practice was not even *the right kind* of justification for our practices as we understand them Strawson said the incompatibilists would counter.[112] In response to this, it's important that a view on which we need the responsibility system because it's necessary to sustain the good life need not be read in a shallow means/end or utilitarian way. There are other ways we can understand the practice and the responsibility system could be an integral component living a dignified life for example, partly constitutive of a virtuous life where 'virtuous life' is not reducible or able to be captured in terms of consequentialism or contractualism. So an argument that we need the responsibility system need not be read in such a way as would give rise to Strawson's worry.

4.12 The practical necessity of moral responsibility and the reactive attitudes

In this section I will firstly lay out the argument for the necessity of moral responsibility and then in the second half turn to dealing with criticisms of it, both existing and anticipated. In particular, as I have already mentioned above, it will be important to address the worry voiced by P. F. Strawson in *Freedom and Resentment* where any attempt to give a practical means end justification of the responsibility system and reactive attitudes is 'not even the right kind' of justification. This is presumably because people are working with a basic desert conception of moral re-

[112] Strawson (1962)

sponsibility, the sense of responsibility which Pereboom says is the 'one at issue' in the free will debate. The basic desert conception entails that someone can be held responsible simply in virtue of having knowingly done something wrong and not in virtue of the anticipated effects of the attribution of responsibility or any other reason of that kind. So the worry is that working in a compatibilist revisionist framework I will not end up doing justice to the basic desert conception of responsibility that is at play in our everyday understanding. I will address this worry after laying out what I take to be the general justification for the responsibility system.

From childhood development and continuing throughout a person's life, the deployment of the reactive attitudes and the (now revised) concept of moral responsibility is necessary for the moral development and functioning of human beings. It's necessary for them to develop moral understanding in general, empathy and sensibility in general. It's necessary for us to be able to live together safely in communities and work together. This point is discussed less than it should be in the literature. Without attributions of responsibility and the reactive attitudes as we know them it is difficult to overstate how much of human life as we know it would cease to exist. I don't just mean by this that the reactive attitudes themselves would cease to exist but in addition that the norms and patterns of behaviour that those attitudes helped foster and sustain would be in jeopardy as well. Perhaps it could be said in reply that other forms of moral censure could be deployed instead that would do the

same job as the reactive attitudes. I think that when the details are fleshed out as to what such an alternative set of attitudes would look like we would end up with something practically very close to what I am recommending here.[113] What is needed is a set of attitudes and related concept of moral responsibility that can play a certain kind of role, namely help shape the kind of agents and community we have independent reasons for wanting to live with and within here. To the extent that any attitudes and practices do this successfully, they will be converging on the revisionist theory I am advocating here. In summary, given the consequences of abandoning (were it possible) the responsibility system and the plausible premise that we do not want to live in such a world, we have the makings of an argument for the practical necessity of the responsibility system. Surely we want to live in a society and culture where human sensibility is conditioned by a responsibility system.

Is such an argument circular? Perhaps it might be said that the desirability of the responsibility system I am using in the above argument is based on the fact that we are making judgements about its benefits from a culture with a responsibility system and that this is not a neutral place to make the argument from. One idea might be that if we were living in a society without reactive attitudes then we may not feel the pull of such

[113]Even supposing that Strawson's objective attitude were an option combined with cool non-affective censure and praise, it would still need to be considered whether it was more or less effective than a revised system of the reactive attitudes and concept of moral responsibility as we commonly understand them. Working on the assumption that the objective attitude is not possible here, I leave these worries aside.

an argument at all. In response to this it shouldn't be controversial to say that the general aims and values I am using to underwrite the necessity of the responsibility system here are meant to be so general that we would also endorse them in a possible world where we didn't maintain a responsibility system in anything like its current form. At its core the responsibility system makes possible a safer world where individual agents can pursue their goals more easily then they would be able to otherwise. This type argument can gain traction with all kinds of sentient beings who are able to plan and carry out intentional action over time. It should be clear that these assumptions are not in any way internal to the perspective of the responsibility system. It also makes possible a more just world, where standards of justice can be underwritten independently of any prior commitment to a responsibility system. Many options are on the table here, but to take one powerful recent example, you could generate a system of political norms via a contract theory in the way John Rawls or Ronald Dworkin have and then be in a position to justify the responsibility system (and for that matter the criminal justice system) on the basis of your foundational political theory. In both cases mentioned here there is no circularity.

It might be said that if the argument for the necessity of the responsibility system goes through in the way I've outlined, it proves too much. Perhaps the argument is powerful enough to justify not only a revisionist compatibilist responsibility system but potentially the existing (and arguably) incompatibilist responsibility system as it stands. Given that our

current practice (committed to incompatibilist content as I have argued) plausibly has some of, if not all of the desired effects that I've based the argument on don't we also have an argument to justify existing practice, to justify the status quo? If sticking with our current system of reactive attitudes and the notions of basic desert (in the incompatibilist sense) they would entail delivered the same desirable effects why not stick with that? In fact it looks like as long as responsibility concepts are successful at generating the desired effects in the way I've outlined as necessary to maintain moral understanding, sensibility and consequent behaviour we can justify both compatibilist and incompatibilist concepts equally well. What does this mean for my overall aim as a prescriptive compatibilist? In response to this worry I want to point out that whether a particular system of responsibility concepts and practices delivers on the overall aim of fostering the desired moral outcomes is not the only factor when deciding which system we should endorse. Other factors must include parsimony and truthfulness for example. So if, unlike their compatibilist counterparts, incompatibilist systems presuppose concepts not realisable in the actual world, then all things considered we will have an argument for the prescriptive compatibilist responsibility system as preferable to a justification for our practices as they stand on the assumption that diagnostic incompatibilism was true.

4.13 Further worries about the revisionist project

In comments on an earlier version of this chapter Helen Steward has pressed three worries: firstly, the issue of prescribing for the causally determined about *action* itself, secondly, issues related to the 'ought implies can' principle and lastly on the claims I have made about inferential role preservation. I respond to these comments below.

Firstly, Steward asks about whether a prescriptive strategy could work for agency incompatibilism as opposed to the more standard incompatibilism of free will that I discuss in this thesis. Steward holds the view that agency itself is incompatible with determinism and moral responsibility is also incompatible with it because responsibility requires agency. Hence could it be argued (as I attempted to do concerning free will and responsibility here) that we *should* adopt a compatibilist conception of agency *if* our current concept is incompatibilist at the moment. Without going into a detailed conceptual analysis of our concept of agency, I would stand by an analogous treatment of this issue and claim that in principle the answer depends on the amount of inferential role that could be preserved if we were to attempt that. The nature of action is of course a philosophically controversial topic but it might be the case that all sides (including the incompatibilist) could agree that there was a lot of inferential role left intact once incompatibilist content was revised away, regardless of your initial position on whether that incompatibilist content was necessary for the concept as it stood in ordinary language and practice.

For example, Donald Davidson's core ideas about action being something someone does that is 'intentional under some description' are plausibly captured in both compatibilist and incompatibilist models and central distinctions between passivity and activity are likewise renderable within the compatibilist camp. This can still be true *even if* as a matter of fact you were an agency incompatibilist about our current concept. As I have already said, claims of essentialism in these revisionary contexts seem to amount to verbal disputes as opposed to substantive ones. The substantive issue is about the degree of preservation of inferential role. I therefore see no principled reason (without undertaking an analysis of action) why I couldn't argue in an analogous fashion for a prescriptive compatibilism about the concept of agency itself.

Steward's second point concerns issues related to 'ought implies can'. If 'ought' does imply 'can' then my claim that we should adopt a revised compatibilist conception of moral responsibility entails that we can do so (in some sense of 'can'). Steward comments, "But how do we know that we can do so? Perhaps incompatibilist thinking is part of our cognitive architecture."[114] In response I wish to re-emphasise the distinction between the metaphysical or what List would call microphysical sense of possibility and the agential sense of possibility pertinent to claims like, 'I can walk the dog'. My claim is that we *should* be working with the agential sense and not the microphysical sense when we make ability claims for agents, even though I stand by my *diagnostic* leeway incompatibilism

[114] Steward, PhD examiners report, (2014)

about everyday 'could' claims pre-revision. Hence I can consistently deny we have certain abilities (as the relevant concepts of those abilities are diagnosed as requiring incompatibilist content) *but maintain that we can do what we should do in the agential sense, which we should adopt.* But can we adopt the agentive sense of 'should'? Yes, we can adopt revised compatibilist conceptions of these things because that is *within our power as agents in the agential reference frame*, as opposed to flying to the next galaxy for example, or thinking as God might think which both clearly are not. All of this is consistent with the fact that determinism entails only one microphysical future and the fact that (given supervenience) some people so determined will never as a matter of fact adopt the agentive sense of could etc. It may well be right that modalities or alternative possibilities are part of our cognitive architecture, that a 'leeway' conception is in some sense necessary for thinking as we do but I don't think it's true that metaphysical possibility (that would require incompatibilism) has to be a necessary part of that leeway conception. Steward's point raises an important test though for incompatibilists who believe determinism to be true (or that it might well be true). What sense can they make of 'should' claims when they also endorse 'ought implies can' if they are not to embrace revisionism as I have? Interestingly, on this issue Pereboom explores the options for an 'axiological recommendation' reading of 'ought' claims that is consistent with determinism and linked with responsibility in a forward looking sense.[115] I don't think we need to make

[115] See Pereboom (2014: 138-146)

that move given the availability of the agentive sense of possibility already discussed but it is worth mentioning as it is another possible response to this specific worry.

Thirdly, Steward says the following on inferential role:

> …the criterion you suggest for deciding whether a concept can survive substantial changes in various beliefs related to it is whether most of its inferential role is preserved in the new context. But how do we decide what constitutes sameness of inferential role? To take an example, suppose we all suddenly realised that Berkeleian idealism was true. In that case, one might think, the word 'table' no longer really means quite what it did before. But one might think to suggest that most of its inferential role had been preserved, despite the change in a very fundamental belief about the nature of tables. But has it? Can I e.g. infer that if there's a table in this room, there's a piece of furniture in this room? Only if the content of the concept 'furniture' also changes accordingly, so that 'furniture' also now means something different from what it meant before, something compatible with the idealist nature of tables. Arguably, therefore, the inferential role of a concept like 'table' has to change *massively* to accommodate the shift to Berkeleian idealism - and then we'd have to say it wasn't really the same concept at all. And one could apply a like argument in the case of moral responsibility. E.g. 'if S is morally responsible for result R, then (other relevant conditions being satisfied) it is just to punish S for result R' can be preserved, one might say, only if the content of 'just' and 'punish' are altered accordingly. And this might threaten to rob the inferential role criterion of any power, mightn't it?[116]

In reply to this third worry, I will first outline my response to the purported problem of the inferential role of 'table' if Berkeleian Idealism turned out to be true and then I will consider if an analogous reply can work for the concept of moral responsibility. I would claim that in the Berkeleian scenario Steward outlines, there would likely have to be con-

[116] Steward, PhD examiners report, (2014)

cept revision at a much higher level of abstraction, namely for concepts like 'material object' and 'space' and 'dimension' and so on. Furthermore, if it *was* possible to make suitable changes at *that* level of abstraction (and this admittedly might include some instances of concept rejection) then many other concepts at lower levels, like 'furniture' and 'table' could in principle maintain (most of) their entailment relations with each other as we would want. For example, if the more abstract concept 'material object' as well as notions like 'spacial dimension' are suitably revised, then it seems as if the inferential role of the *instances* of those more general concepts would allow us to infer that 'there is a piece of furniture in the room' from 'there is a table in the room.' That much seems true. The sticking point might be whether it was feasible to modify and revise the more abstract concepts along the lines I have suggested above in the Berkeleian scenario. On the other hand, as a second suggestion here, even *if* some of the more abstract concepts of 'space time' and notions of 'matter' had to be rejected altogether, I would still argue that once new concepts had been adopted at a suitably high level of abstraction in their place, there is in principle no barrier to all of the less abstract concepts (like 'table' and 'furniture') being (all together) modified in accordance with those changes. In that case it may well be that the vast majority of inferences involving concepts at the lower levels of abstraction can carry on much as before. Will something like this reply speak to the analogous worry about responsibility and the link with the concept of punishment, as Steward outlines it above? In principle I don't see why not though I

also want to claim here that there might be a disanalogy between these cases. In short, it might be that there just isn't a suitable compatibilist reading of 'punish' because that notion essentially relies on an incompatibilist model of agency to make sense of its retributive content. Without arguing for this claim, if that *was* the case then it might well be true that we could no longer infer 'it's just to punish S' (all other conditions satisfied) from 'S is morally responsible' (in the revised sense). But that might be ok just because it's no longer just to punish anyone anyway. In that case, as long as enough of the content of 'moral responsibility' is preserved, e.g. S wasn't coerced, S knew what they were doing, S didn't (in the agentive sense of possibility) *have* to do that etc, then there won't necessarily be a problem here for *other* kinds of entailments of the concept of responsibility post revision. That is to say, enough of the inferential role can in principle still be preserved even if we have to junk some concepts like 'punishment.' As for the other concept Steward mentions, 'justice', I would claim it is possible to rely on a notion of justice that is not contingent on the metaphysical and moral issues at the heart of the free will problem. One suggestion here would be to make use of a Rawlsian notion of justice which would allow us to make inferences in these kinds of contexts regardless of the outcomes to issues in the free debate.

4.14 Vargas' six worries about 'moral influence' theories

In *Building Better Beings*, in the section on 'Justifying the Practice' Manuel Vargas considers six worries about the approach he and I wish to take here. Six criticisms of the core idea that responsibility practices can ultimately be justified in the way I have contented - because of their desirable effects. I will conclude this thesis by going through these worries in turn. Vargas claims these are the key objections to moral influence theories developed during the last fifty years. My argument has been that we need the responsibility system as it is necessary to maintain the type of moral practice, i.e. behaviours and sensibility in people for community to flourish. As such I am squarely in the sights of these worries and they must be addressed.

The first worry is that moral influence (MI) theories are not able to make the appropriate distinctions between different types of agents. It looks like this might be a problem if both intuitively and non-intuitively responsible agents are equally susceptible to being influenced by the reactive attitudes. More specifically, how do we explain the difference between the standard we hold adult humans to in comparison with young children and animals. The second worry is that it's hard to see how MI theories are able to distinguish between moral influence (as distinctively moral) and influence in general that gets the same results as moral influence. These first two worries are related. The third worry is that the MI account seems to equate being responsible with considerations of

whether it would be rational or appropriate to hold somebody responsible (depending on whether they can/can't be influenced in the right kinds of ways). We commonly make a distinction between those things. Fourthly, Vargas worries that MI accounts fail to accurately describe *how* we hold people morally responsible. It is consistent with the requirements of a MI account that we never actually feel any of the reactive attitudes required by the account but just behave as if we do in order to secure the desired influence. Does this leave out the vital hostility that is arguably central to blame? The fifth worry is also a worry about misdescribing our practice here. If we praise or blame someone now dead for some action they had done in the past, it would appear that standard moral influence theories require that the attitudes expressed have influence, but this is not possible for the deceased so it must be for others still alive. However, it just looks unnecessary that this must be in place in order for every instance of praising and blaming historic figures to be justified. Surely some instances are justified even if there was no influence at all. Lastly, the sixth worry is that MI theories are tied to some type or another of consequentialism and given that this is controversial ground, the justification of the responsibility system cannot be left meta-ethically contingent in that way. I will now attempt to address these worries in turn.

Vargas thinks that many of these worries can be dealt with by a shift from requiring MI justifications for specific acts of blaming or praising to instead requiring them of general practice. He says:

> There is, I think, a temptation on the part of both moral influence theorists and their critics to think of the justification of moral influence in terms of the efficacy of particular tokenings of praising and blaming, of the practices of expressing those judgments and acting in characteristic ways upon them. So, when, say, Lori criticizes Dan for being overly self-conscious, proponents and critics of moral influence accounts have tended to think that the justificatory force derives from the efficacy of the particular instance of so criticizing or praising in that particular time and circumstance.
>
> That's a view. A more plausible view, however, construes the justification for moral praise and blame as arising not at the level of particular interpersonal interactions, but instead, at the level of a general practice. On the model I propose, the justification arises from the group-level effects of justified norms that are ubiquitously internalized by members of the community and regularly put into practice. This difference—both a scaling back of ambition and elevation of the source of justification from tokens to the in-practice effects of the system of norms of praise and blame as a whole—turns out to dissolve many of the familiar objections to the idea of moral influence.[117]

This so called 'scaling back of ambition' helps us to resolve the six worries about MI theories. With regard to the first worry, that it's difficult to make distinctions between agents, the MI theory itself is not supposed to explain this distinction in terms of influence. At a different level of abstraction an account of reasons-responsive or as Vargas develops 'moral considerations-responsive agency' is meant to do that. The MI theory is what ultimately justifies the responsibility system but it is not supposed to explain all the dynamics *within* the practice it justifies. It is not, in other words a theory of moral responsibility itself. To ask it to be that is to look for justification at the wrong level of abstraction here. I have of course not developed a detailed account of responsible agency in this

[117] Vargas (2013: 172-3)

thesis but am sympathetic with some kind of reasons responsive position. The main point here is to clear up the worry based on looking for a justification at the wrong level. When we look for it at the right level, we should hopefully be able to draw the distinctions the worry outlined in the correct place and in consequence be equipped to distinguish satisfactorily between animals and humans, between children, adults and psychopaths etc. The second worry is dealt with in a similar way to the first. In distinguishing between moral and non moral influence, we need to look at the theory of responsibility we are working with itself and not to what ultimately justifies the responsibility system (MI). So again, a suitable reasons-responsive account will hopefully tell us what the conditions are for holding someone morally responsible but of course these will not delimit other forms of influence. That is why those other forms of influence will not be candidates for holding someone responsible. Vargas says:

> …on this account, the appropriateness of praise and blame is parasitic on the truth of the judgment that the target of praise and blame is a responsible agent. And, as we have seen, that is given by a theory of responsible agency and not a theory of the justification of the responsibility norms.
>
> In contrast, other forms of influencing the behavior of agents have no such requirement that the agent be a responsible agent, and indeed no such supposition ordinarily built into them. In influencing a household pet, there is (ordinarily) no judgment that the pet is a responsible agent. Hence, the form of regard expressed in distinctively moral praise and blame is not present. So, even if some of the practices of moral influence are superficially indistinguishable from non-

> moral influence, the underlying attitudes and judgments are distinct.[118]

What about the third worry, that MI accounts appear to claim that someone is only actually responsible when we should (in terms of expected influence) hold them responsible? From the Vargas quotes above it should be relatively clear as to the way to reply here. The appropriate way to answer the question 'is an agent morally responsible?' is with reference to the particular theory of moral responsibility (i.e. reasons-responsiveness) you have adopted and not the ultimate justification of us developing a responsibility system. Moreover it is consistent with the MI theory here that the particular judgements of the responsibility system (i.e. that Fred *is* morally responsible) can run coherently side by side with us having a reason not to deploy reactive attitudes in any particular case. Considerations of prudence and benevolence can still happily trump here as they can in the responsibility system as it currently stands.

As for the fourth objection, the fact that blame/praise like behaviour *could* be justified (if it were possible to produce it) by the same ultimate MI argument is not really a criticism at all. As long as our actual practices are justified (ultimately in terms of their effects), why should we have to worry that other practices could be as well? The fact that they could doesn't undermine the justification put forward here. More critically, there will be good reasons contingent on human psychology why our responsibility practices (with their attendant affective quality) are in

[118] Vargas (2013: 189)

place rather than merely the outward appearances of them. It's not clear for example that the latter is sustainable or coherent.

With respect to the fifth objection, the particular theory of responsibility we adopt itself decides whether figures in the past were actually responsible (not the MI theory). Given that we've already seen that particular tokens of praise and blame don't all need to be justified in a forward looking way anyway, we can reply here that as long as these instances (for example blaming the dead) are part of a general practice that is well justified internal to the responsibility system - and a good justification would be the fact that these instances would reinforce the appropriate sensitivities, judgements and norms) - there is no problem.

Vargas argues that the sixth objection doesn't apply as accounts such as his are 'explicitly modular' and not committed to any particular account of normative ethics. Earlier in this chapter I argued how a MI theory could find justification within a meta-ethical framework that was explicitly virtue ethics centred as opposed to consequentialist, contractarian or deontological. Once again, keeping the different types of justification for different parts of an overall theory clearly distinct should help dispel this worry.

In summary, even though I have not (unlike Vargas) developed a detailed theory of moral responsibility here, these general worries about MI justification needed to be discussed because MI is ultimately what justifies the responsibility system.

4.15 Final Note

Finally, it's always a good strategy to turn the challenge around and ask the responsibility sceptic how they think we *would and should* manage without the responsibility system. The effects of ceasing to hold people responsible, both children and adults would likely be significant. It is already well documented that children who are raised without being held accountable for their behaviour often turn out to be antisocial and have problems which makes it difficult for them to sustain relationships with others. Whether or not there was thought to be an incompatibilist condition on the folk concept of responsibility, it is plain that the price of giving up on the responsibility system would be too high (revised or otherwise). We *must* continue with it, albeit in revised form and this provides justification enough.

The incompatibilist sceptic may reply to the normative argument that they agree we do need a system of censure and moral feedback for exactly the above reason but that we shouldn't call it a system of moral responsibility. Presumably this is again for the reason that they take a necessary condition on responsibility to be unrealisable. In this case however, the justification for the responsibility system is partly tied to the normative argument and also to the fact that we can revise our concept if enough of the work of the concept is preserved without the incompatibilist condition. So, in conclusion it's important to run the two arguments together. It's the combination of the possibility of conceptual revision

(because sufficient inferential role is preserved without the incompatibilist condition) *and* the fact that we need the responsibility system, that together provide the best argument against the responsibility sceptic. We need the revisionist argument about the preservation of majority inferential role of the concept to sure up the claim that it's 'responsibility' we're talking about and not something else and we need the normative argument to respond to the 'throwing up of hands' request for the all things considered rationale of the responsibility system. I believe that these approaches together provide a sufficient and harmonious defence of prescriptive compatibilism.

Bibliography

Alvarez, Maria. (2009). "Actions, Thought-Experiments, and the 'Principle of Alternate Possibilities'," *Australian Journal of Philosophy* 87, pp. 61–81.

Austin, J. L. (1961). "Ifs and Cans" in *Philosophical Papers*: London: OUP, pp. 153-180.

Ayer, Alfred J. (1954). "Freedom and Necessity," in A. J. Ayer, *Philosophical Essays*. London: Macmillan, pp. 271–84.

———— (1984). *Freedom and Morality and Other Essays,* Oxford: Clarendon Press

Balaguer, Mark. (2009). Free Will as an Open Scientific Problem. Cambridge, MA: MIT Press.

Björnsson, Gunnar, and Derk Pereboom. (2014). "Free Will Skepticism and Bypassing," in W. Sinnott-Armstrong, ed., *Moral Psychology Vol. 4*. Cambridge, MA: MIT Press, pp. 27–35.

Bok, Hilary. (1998). *Freedom and Responsibility*. Princeton: Princeton University Press.

Capes, Justin. (2010). "The W-Defense," *Philosophical Studies* 150, pp. 61–77.

———— (2012). "Action, responsibility and the ability to do otherwise," *Philosophical Studies* 158, (1): pp. 1-15.

Chisholm, Roderick. (1964). "Human Freedom and the Self," The Lindley Lecture, Department of Philosophy, University of Kansas.

Clarke, Randolph. (2003). *Libertarian Theories of Free Will*. New York: Oxford University Press.

———— (2008). "Incompatibilist (Nondeterministic) Theories of Free Will," in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy.*

———— (2009). "Dispositions, Abilities to Act, and Free Will: The New Dispositionalism," *Mind* 118, pp. 323–51.

Dennett, Daniel. (1984). *Elbow Room*. Cambridge, MA: MIT Press.

———— (2003). *Freedom Evolves*. New York: Viking.

Fara, Michael. (2008). "Masked Abilities and Compatibilism," *Mind* 117, pp. 843–65.

Fischer, John Martin. (1994). *The Metaphysics of Free Will*. Oxford, Blackwell Publishers.

———— (2003). "'Ought-Implies-Can,' Causal Determinism, and Moral Responsibility," *Analysis* 63, pp. 244–50.

———— (2004). "Responsibility and Manipulation," *The Journal of Ethics* 8, pp. 145–77.

———— (2006). *My Way*. Oxford: Oxford University Press.

———— (2009). *Our Stories*. New York: Oxford University Press.

———— (2010). "The Frankfurt Cases: The Moral of the Stories," *Philosophical Review* 119, pp. 315–36.

Fischer, John Martin, Robert Kane, Derk Pereboom, Manuel Vargas (2007). *Four Views on Free Will*. Oxford: Blackwell.

Fischer, John Martin, and Mark Ravizza (1998). Responsibility and Control: A Theory of Moral Responsibility. Cambridge: Cambridge University Press.

Frankfurt, Harry. (1969). "Alternate Possibilities and Moral Responsibility," *Journal of Philosophy* 66, pp. 829–39.

———— (1971). "Freedom of the Will and the Concept of a Person," *Journal of Philosophy* 68, pp. 5–20.

Franklin, Christopher. (2011). "Neo-Frankfurtians and Buffer Cases: The New Challenge to the Principle of Alternative Possibilities," *Philosophical Studies* 152, pp. 189–207.

Ginet, Carl. (1966). "Might we have no Choice?" in Keith Lehrer, ed., *Freedom and Determinism*. New York: Random House.

———— (1990). *On Action*. Cambridge: Cambridge University Press.

———— (1996). "In Defense of the Principle of Alternative Possibilities: Why I Don't Find Frankfurt's Arguments Convincing," *Philosophical Perspectives* 10, pp. 403–17.

———— (2002). "Review of *Living without Free Will*," *Journal of Ethics* 6, pp. 305–9.

Haas, Daniel. (2012). "In Defense of Hard-line replies to the Multiple-case Manipulation Argument," *Philosophical Studies* 163, 797–811.

Haji, Ishtiyaque. (1998). *Moral Accountability*. New York: Oxford University Press.

———— (2004). "Active Control, Agent Causation, and Free Action," *Philosophical Explorations* 7, pp. 131–48.

———— (2012). *Reason's Debt to Freedom*. New York: Oxford University Press.

Haji, Ishtiyaque, and S. Cuypers. (2006). "Hard- and Soft-Line Responses to Pereboom's Four-Case Manipulation Argument," *Acta Analytica* 21, pp. 19–35.

Hunt, David. (2000). "Moral Responsibility and Unavoidable Action," *Philosophical Studies* 97, pp. 195–227.

———— (2005). "Moral Responsibility and Buffered Alternatives," *Midwest Studies in Philosophy* 29, pp. 126–45.

Hunt, David and Shabo, Seth. (2013) "Frankfurt cases and the (in)significance of timing: a defense of the buffering strategy" *Philosophical Studies* 164: 599–622.

Kane, Robert. (1985). *Free Will and Values*. Albany: SUNY Press.

———— (1996). *The Significance of Free Will.* New York: Oxford University Press.

———— (2000). "The Dual Regress of Free Will and the Role of Alternative Possibilities," *Philosophical Perspectives* 14, pp. 57–80.

Kearns, Stephen. (2012). "Aborting the Zygote Argument," *Philosophical Studies* 160, pp. 379–89.

Kelly, Erin (2009). "Criminal Justice without Retribution," *Journal of Philosophy* 106, pp. 440–62.

Larvor, B. (2010). "Frankfurt Counter-Example Defused" *Analysis*, Vol 70, No. 3.

Lehrer, Keith. (1968). "Cans without Ifs," *Analysis* 29, pp. 29–32.

Lewis, David. (1981). "Are we free to break the laws?" *Theoria* 47, pp. 113-121.

List, Christian. (2014). "Free Will, Determinism, and the Possibility of Doing Otherwise", *Nous* 48:1, pp. 156–178

Lowe, E. Jonathan. (2008). Personal Agency: The Metaphysics of Mind and Action. Oxford: Oxford University Press.

McKenna, Michael. (2003). "Robustness, Control, and the Demand for Morally Significant Alternatives," in Michael McKenna and David Widerker, eds., *Freedom, Responsibility, and Agency: Essays on the Importance of Alternative Possibilities*. Aldershot, UK: Ashgate, pp. 201–16.

———— (2005). "Where Frankfurt and Strawson Meet," *Midwest Studies in Philosophy* 29, pp. 163–80.

———— (2008). "A Hard-Line Reply to Pereboom's Four-Case Argument," *Philosophy and Phenomenological Research* 77, pp. 142–59.

———— (2008). "Saying Good-Bye to the Direct Argument in the Right Way," *Philosophical Review* 117, pp. 349–83.

———— (2008). "Frankfurt's Argument against the Principle of Alternative Possibilities: Looking Beyond the Examples," *Noûs* 42, pp. 770–93.

Mele, Alfred. (1995). *Autonomous Agents*. New York: Oxford University Press.

———— (2005). "Libertarianism, Luck, and Control," *Pacific Philosophical Quarterly* 86, pp. 381–407.

———— (2005). "A Critique of Pereboom's 'Four-Case' Argument for Incompatibilism," *Analysis* 65, pp. 75–80.

———— (2006). *Free Will and Luck*. New York: Oxford University Press.

Mele, Alfred, and David Robb. (1998). "Rescuing Frankfurt-Style Cases," *Philosophical Review* 107, pp. 97–112.

———— (2003). "Bbs, Magnets and Seesaws: The Metaphysics of Frankfurt-style Cases" in Widerker and McKenna, (Eds.), pp. 132-136.

Moya, Carlos. (2006). *Moral Responsibility: The Ways of Skepticism*. Oxford, Oxford University Press, pp. 195–210.

———— (2011). "On the Very Idea of a Robust Alternative," *Critica* 43, pp. 3–26.

Nelkin, Dana. (2004). "Deliberative Alternatives," *Philosophical Topics* 32, pp. 215–40.

———— (2008). "Responsibility and Rational Abilities: Defending an Asymmetrical View," in *Pacific Philosophical Quarterly* 89, pp. 497–515.

———— (2011). Making Sense of Freedom and Responsibility. Oxford: Oxford University Press.

Nichols, Shaun, and Joshua Knobe. (2007). "Moral Responsibility and Determinism: The Cognitive Science of Folk Intuitions," *Nous* 41, pp. 663–85.

O'Connor, Timothy. (1995). "Agent Causation," in Timothy O'Connor, ed., *Agents, Causes, and Events*. New York: Oxford University Press, pp. 170–200.

———— (2000). *Persons and Causes*. New York: Oxford University Press.

Otsuka, Michael. (1998). "Incompatibilism and the Avoidability of Blame," *Ethics* 108, pp. 685–701.

Palmer, David. (2011). "Pereboom on the Frankfurt Cases," *Philosophical Studies* 153, pp. 261–72.

Pereboom, Derk. (1995). "Determinism *Al Dente*," *Noûs* 29, pp. 21–45.

———— (2000). "Alternative Possibilities and Causal Histories," *Philosophical Perspectives* 14, pp. 119–37.

———— (2001). *Living without Free Will*. Cambridge: Cambridge University Press.

———— (2003). "Source Incompatibilism and Alternative Possibilities," in M. McKenna and D. Widerker, eds., *Freedom, Responsibility, and Agency: Essays on the Importance of Alternative Possibilities*. Aldershot, UK: Ashgate Press, pp. 185–99.

———— (2005). "Defending Hard Incompatibilism," *Midwest Studies in Philosophy* 29, pp. 228–47.

———— (2008). "A Hard-line reply to the Multiple-Case Manipulation Argument," *Philosophy and Phenomenological Research* 77: pp. 160–70.

———— (2009). "Further Thoughts about a Frankfurt-Style Argument," *Philosophical Explorations* 12, pp. 109–18.

———— (2009). "Hard Incompatibilism and its Rivals," *Philosophical Studies* 144, pp. 21–33.

———— (2012). "Frankfurt Examples, Derivative Responsibility, and the Timing Objection," *Philosophical Issues* 22, pp. 298–315.

Russell, Paul. (1992). "Strawson's Way of Naturalizing Responsibility," *Ethics* 102, pp. 287–302.

Sartorio, Carolina (2011). "Actuality and Responsibility," *Mind* 120, pp. 1071–97.

———— (2013). "Making a Difference in a Deterministic World," *Philosophical Review* 122, pp. 189–214.

Shabo, Seth. (2010). "The Fate of the Direct Argument and the Case for Incompatibilism," *Philosophical Studies* 150, pp. 405–24.

———— (2012). "Incompatibilism and Personal Relationships: Another Look at Strawson's Objective Attitude," *Australasian Journal of Philosophy* 90, pp. 131–47.

Smilansky, Saul. (2000). *Free Will and Illusion*. New York: Oxford University Press.

Steward, Helen. (2009). "Fairness, Agency, and the Flicker of Freedom," *Noûs* 43, pp. 64–93.

———— (2012). *A Metaphysics for Freedom*, Oxford: Oxford University Press.

Strawson, Galen (1986). *Freedom and Belief*. Oxford: Oxford University Press.

Strawson, Peter F. (1962). "Freedom and Resentment," *Proceedings of the British Academy* 48, pp. 1–25.

Todd, Patrick. (2011). "A New Approach to Manipulation Arguments," *Philosophical Studies* 152, pp. 127–33.

———— (2012). “Manipulation and Moral Standing: An Argument for Incompatibilism,” *Philosophical Imprint* 12, pp. 1–18.

———— (2013). “Defending (a Modified Version of the) Zygote Argument,” *Philosophical Studies* 164, pp. 189–203.

van Inwagen, Peter. (1975). “The Incompatibility of Free Will and Determinism,” *Philosophical Studies* 27, pp. 185–99.

———— (1983). *An Essay on Free Will*. Oxford: Oxford University Press.

Vargas, Manuel. (2007). “Revisionism” and “Response to Fischer, Kane, and Pereboom,” in J. Fischer, R. Kane, D. Pereboom, M. Vargas, *Four Views on Free Will*, Oxford: Blackwell Publishers, pp. 126–65; 204–19.

———— (2013). *Building Better Beings, A Theory of Moral Responsibility*, Oxford: Oxford University Press.

Vihvelin, Kadri. (2004). “Free Will Demystified: A Dispositional Account,” *Philosophical Topics* 32, pp. 427–50.

_____ (2011). "Arguments for Incompatibilism", The Stanford Encyclopedia of Philosophy, Edward N. Zalta (ed.)

———— (2013) *Causes, Laws, & Free Will, Why Determinism Doesn't Matter*, New York: Oxford University Press.

Wallace, R. Jay. (1994). *Responsibility and the Moral Sentiments*. Cambridge: Harvard University Press.

———— (2011). “Dispassionate Opprobrium: On Blame and the Reactive Sentiments,” in R. J. Wallace, R. Kumar, and S. Freeman, eds., *Reasons and Recognition. Essays on the Philosophy of T. M. Scanlon*. New York: Oxford University Press, pp. 348–72.

Watson, Gary. (1975). "Free Agency," *Journal of Philosophy* 72, pp. 205–20.

———— (1987). "Responsibility and the Limits of Evil," in Ferdinand Schoeman, ed., *Responsibility, Character, and the Emotions*. Cambridge: Cambridge University Press, pp. 256–86.

———— (1996). "Two Faces of Responsibility," *Philosophical Topics* 24, pp. 227–48.

Widerker, David. (1995). "Libertarianism and Frankfurt's Attack on the Principle of Alternative Possibilities," *The Philosophical Review* 104, pp. 247–61.

———— (2000). "Frankfurt's Attack on Alternative Possibilities: A Further Look," *Philosophical Perspectives* 14, 181–201.

_____ (2003). "Blameworthiness and Frankfurt's Argument Against the Principle of Alternative Possibilities" in Widerker and McKenna, (Eds.)

———— (2006). "Libertarianism and the Philosophical Significance of Frankfurt Scenarios," *Journal of Philosophy* 103, pp. 163–87.

Wolf, Susan. (1980). "Asymmetrical Freedom," *Journal of Philosophy* 77, pp. 151–66.

———— (1990). *Freedom within Reason*. Oxford: Oxford University Press.